TITLE OF THE INVENTION

SYSTEM, METHOD, DEVICE, AND COMPUTER PROGRAM PRODUCT FOR EXTRACTION, GATHERING, MANIPULATION, AND ANALYSIS OF PEAK DATA FROM AN AUTOMATED SEQUENCER

REFERENCE TO COMPUTER PROGRAM LISTING APPENDIX

Filed herewith in triplicate (labeled Copy 1.1, Copy 1.2, and Copy 1.3, respectively) is a computer program listing appendix on a compact disc read only memory (CD-ROM). The entire contents of the computer program listing appendix is incorporated herein by reference.

Each of the three copies of the computer program listing appendix were created on July 1, 2003, and each includes the following files:

Name	Size	Type	Modified
Admin	49KB	Text	7/1/02
CAppFile.cp	7KB	CP file	2/5/02
CAppFile.h	3KB	H file	3/30/01
CAppleScript.cp	8KB	CP file	1/29/01
CAppleScript.h	2KB	H file	1/29/01
CGel.cp	6KB	CP file	8/16/01
CGel.h	4KB	H file	1/29/01
CISEApeaksApp.cp	19KB	CP file	2/7/02
CISEApeaksApp.h	2KB	H file	1/29/01
CISEApeaksApp.ppob	0KB	PPOB file	2/4/02

CISEApeaksApp.rsrc	0KB	RSRC file	6/19/01
CISEApp.cp	11KB	CP file	4/20/01
CISEApp.h	2KB	H file	3/27/01
Common.ppob	0KB	PPOB file	11/10/00
Constants.h	5KB	H file	6/20/02
Constants	24B	TEXT	7/1/02
COutPutDoc.cp	5KB	CP file	3/27/01
COutPutDoc.h	3KB	H file	1/29/01
CPictPlaces.cp	8KB	CP file	8/10/01
CPictPlaces.h	3KB	H file	1/29/01
CPrefFile.cp	4KB	CP file	3/1/01
CPrefFile.h	1KB	H file	1/29/01
CRun.cp	34KB	CP file	2/7/02
CRun.h	3KB	H file	1/29/01
CTextDocument.cp	9KB	CP file	1/29/01
CTextDocument.h	1KB	H file	1/29/01
CTextView.cp	2KB	CP file	1/29/01
CTextView.h	2KB	H file	1/29/01
DataAnalyser	237KB	TEXT	7/1/02
DataFormater	83KB	TEXT	7/1/02
DataParameter	55KB	TEXT	7/1/02
DataUtility	66KB	TEXT	7/1/02
DEApp.ppob	0KB	PPOB file	12/22/00
DEApp.rsrc	0KB	RSRC file	6/19/01
DSApp.ppob	0KB	PPOB file	10/19/00

DSApp.rsrc	0KB	RSRC file	6/19/01
FrmAbout	22KB	TEXT	7/1/02
FrmDFPref	22KB	TEXT	7/1/02
FrmfileInfo	21KB	TEXT	7/1/02
FrmLicence	23KB	TEXT	7/1/02
FrmMainMenu	24KB	TEXT	7/1/02
FrmNewFile	21KB	TEXT	7/1/02
FrmPreferences	25KB	TEXT	7/1/02
FrmYesNo	20KB	TEXT	7/1/02
GSExportscript	25KB	TEXT	7/1/02
GSPeak.h	2KB	H file	1/29/01
GSProfile.cp	3KB	CP file	1/29/01
GSProfile.h	2KB	H file	1/29/01
GUI_Macros	76KB	TEXT	7/1/02
ISEApp.ppob	0KB	PPOB file	3/27/01
ISEApp.rsrc	0KB	RSRC file	6/19/01
ISPeak.h	2KB	H file	1/29/01
ISProfile.cp	15KB	CP file	1/29/01
ISProfile.h	3KB	H file	1/29/01
MyFileUtilities.cp	18KB	CP file	8/10/01
MyFileUtilities.h	2KB	H file	1/29/01
MyUtilities.cp	5KB	CP file	2/6/02
MyUtilities.h	2KB	H file	2/5/02
Peak.h	1KB	H file	2/5/02
PictPlace.h	2KB	H file	1/29/01

PrintingConstants.h	1KB	H file	10/5/00
Profile.cp	21KB	CP file	2/13/02
Profile.cp.old	18KB	CP file	3/1/01
Profile.h	7KB	H file	2/5/02
Thiswbk	21KB	TEXT	7/1/02
Utilities	94KB	TEXT	7/1/02
Well.h	2KB	H file	1/29/01

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a novel computer program product to extract and gather peak information from an automated sequencer or bioinformatics tool into a peak database, and to manipulate and analyze the peak information within the database.

Discussion of the Background

The recent conclusion of several genome sequencing projects, including yeast (*Nature* 1997; 387:suppl. 3-105), human (Venter et al, *Science* 2001; 291:1304-1351), *C. elegans* (*Science* 1998; 282:2012-8), and rice (J. Yu et al., *Science* 2002; 296:79 and S. A. Goff et al., *Science* 2002; 296:92), as well as on-going sequencing efforts, have generated a deluge of DNA sequence information. These DNA sequences encode the basic "message of life." However, cataloguing and probing the vast numbers of genes and the proteins, which they encode, can provide novel insights into cell biology, drug design, and therapeutic strategies.

Accordingly, many new analytical methods have been developed to digest the flood of genome sequence data, including analysis of the transcriptome, proteome and metabolites.

High-throughput analysis of protein targeting and other methods will ascribe new information to proteins and create important links with other large datasets. To fulfill the potential revealed by this genomic information, many challenges have to be met. Among these are indexing and cataloguing of raw DNA and RNA sequence data, identification of genes and the regulation of their expression, characterization of protein activity and protein-protein, protein-ligand, or protein-DNA/RNA interactions.

One such strategy commonly employed is a DNA automatic sequencer. DNA automatic sequencers are used to determine DNA fragment lengths in a wide array of applications: DNA sequencing, microsatellites, Single Nucleotide Polymorphism, Restriction or Amplified Fragment Length Polymorphism, Single Strand Conformation Polymorphism, gene expression quantification and analyses of the immune receptor diversity. All of these applications require access to raw data (peak area and nucleotide length). Raw data being stored in one file per lane, studies rapidly give rise to hundreds of files. However, with the increasing number of samples analyzed, no tool is currently available to allow the extensive and efficient retrieval of this raw data.

Accordingly, there remains a critical need for a novel program for handling the extensive amounts raw data provided by automated sequencers. In addition, there remains a critical need for a novel program for efficient retrieval of this raw data. Moreover, there remains a critical need for a novel program to analyze the extensive amounts of raw data.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a method for high throughput analysis of data. One embodiment being a method of using bioinformatic tools to extract and/or smooth peak data sets according to parameter files and store them in data files. In a further embodiment, particular profiles representing peaks may be created and may be analyzed.

Another object of the invention is a method of building a database.

Another object of the invention is a method of analyzing database by statistical tools. One embodiment being a method of determining prognostic or diagnostic criteria. In a further embodiment, the prognostic and diagnostic criteria are used in the field of physiopathology such as immunotheraphy, cancer treatment, HIV, infectious disease, and/or autoimmune disease.

Another object of the invention is a high-throughput method for analysis of immune repertoires.

Another object of the invention is a high-throughput method for analysis of immune repertoires comprising purifying polynucleotide fragments from biological samples, which contains polynucleotide fragments, synthesizing polynucleotides from purified polynucleotides, and amplifying polynucleotides. In one embodiment, the polynucleotide is amplified by PCR or SDA methods. A further embodiment is related to labeling polynucleotides for detection. Another embodiment is related separating polynucleotides, identifying peaks by determining their position in a separation matrix and area that correspond to labeled polynucleotides.

Another object of the invention is a method comprising:

a) isolating a biological sample;

b) extracting raw data from the biological sample; and

c) compiling the raw data into a database using ISEApeaks.

The objects above can be obtained and performed according the following further objects of the invention.

Another object of the invention is a computer program product, comprising:

a computer storage medium; and

a computer program code mechanism embedded in the computer storage medium for causing a computer to produce an analysis of raw data produced by a separation technique for biomolecules, the computer program code mechanism having

a first computer code device configured to extract a first set of raw data and a second set of raw data from at least one database,

a second computer code device configured to determine a first value from the first set of raw data corresponding to a first characteristic and a second value from the second set of raw data corresponding to the first characteristic, and

a third computer code device configured to store the first value and the second value in a memory. The above object may be performed in combination with these further embodiments. One embodiment is a computer code device configured to retrieve the first value and the second value from the memory and to order the first value and the second value based on a user preference stored in a second memory. One embodiment is a computer code device configured to produce a graphical representation of the first value and the second value as ordered by the fourth computer code device. One embodiment is a computer code device configured to smooth at least one of the first set of raw data and the second set of raw data produced by the separation technique for biomolecules. One embodiment is a computer code device configured to format at least one of the first value and the second value. One embodiment is a computer code device configured to parameterize at least one of the first set of raw data and the second set of raw data. One embodiment is a computer code device, configured to analyze at least one of the first set of raw data and the second set of raw data. One embodiment is a computer code device, configured to export at least one of the first value and the second value.

Another object of the invention is a device, comprising:

at least one extractor configured to extract raw data produced by a separation technique for biomolecules, the extractor including

a processor, and

a computer readable medium encoded with processor readable instructions that, when executed by the processor implement,

an extraction mechanism configured to extract a first set of raw data and a second set of raw data from at least one database,

a characteristic determining mechanism configured to determine a first value from the first set of raw data corresponding to a first characteristic and a second value from the second set of raw data corresponding to the first characteristic, and

an output mechanism configured to store the first value and the second value in a memory.

Another object of the invention is a system, comprising:

a digital repository populated with entries of raw data produced by a separation technique for biomolecules;

a processor; and

a computer readable medium encoded with processor readable instructions that, when executed by the processor implement,

an extraction mechanism configured to extract a first set of raw data and a second set of raw data from the digital repository,

a characteristic determining mechanism configured to determine a first value from the first set of raw data corresponding to a first characteristic and a second value from the second set of raw data corresponding to the first characteristic, and

an output mechanism configured to store the first value and the second value in a memory.

Another object of the invention is a system, comprising:

a digital repository populated with entries of raw data produced by a separation technique for biomolecules;

a processor; and

a computer readable medium encoded with processor readable instructions that, when executed by the processor implement,

an extraction mechanism configured to extract a first set of raw data and a second set of raw data from the digital repository via a network,

a characteristic determining mechanism configured to determine a first value from the first set of raw data corresponding to a first characteristic and a second value from the second set of raw data corresponding to the first characteristic, and

an output mechanism configured to store the first value and the second value in a memory.

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising extracted raw data produced by a separation technique for biomolecules.

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising smoothed raw data, wherein the raw data includes data produced by a separation technique for biomolecules.

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising formatted raw data, wherein the raw data includes data produced by a separation technique for biomolecules.

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising parameterized raw data, wherein the raw data includes data produced by a separation technique for biomolecules..

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising analyzed raw data, wherein the raw data includes data produced by a separation technique for biomolecules.

Another object of the invention is a computer data signal embodied in a carrier wave, said computer data signal comprising exported raw data, wherein the raw data includes data produced by a separation technique for biomolecules.

Another object of the invention is a software package, wherein said software package is embodied by the ISEApeaks package 2.0.1.

The above objects highlight certain aspects of the invention. Additional objects, aspects and embodiments of the invention are found in the following detailed description of the invention.

BRIEF DESCRIPTION OF THE FIGURES

A more complete appreciation of the invention and many of the attendant advantages thereof will be readily obtained as the same becomes better understood by reference to the following Figures in conjunction with the detailed description below.

Figure 1: a computer system 1101.

Figure 2: Typical PBL repertoires of CTR, CM$^-$ and CM$^+$ mice.

cDNA were obtained from PBL of control (CTR), infected without neurological signs (CM$^-$) and infected with CM (CM$^+$) mice. BV-BC CDR3 spectratyping was performed by PCR amplification of cDNA with combinations of BV-specific primers and a BC-specific primer. After run-off with a fluorescent BC-specific primer, PCR products were size separated on an automated DNA sequencer. Sequencing gels were analyzed with the Immunoscope product.

Typical samples are represented. Horizontal axis represents the nucleotide size, centered on a 10 amino acid CDR3 fragment, and vertical axis the fluorescence intensity, in arbitrary units.

Figure 3: Descriminant Analysis separates CM^+ PBL from CM^+ spleen, CM^- PBL and CM^- spleen BV-BC repertoires. Discriminant Analysis was performed on the new set of variables obtained with Principal Components Analysis. Bivariate graphics represent the value of two discriminant functions. Discriminant functions are linear combination of the variables that try to separate groups. F1 to F5 stands for each discriminant function, sorted decreasingly by the associated eigenvalue. Samples (8 CTR PBL, 6 CTR spleen, 10 CM^- PBL, 8 CM^- spleen, 13 CM^+ PBL and 10 CM^+ spleen) are represented on the bivariate plots, but are scarcely visible since DA groups samples very well. Confidence (0.95) ellipses, centered on group centroid, are overlaid. The density normal curve of each group is shown on the diagonal. Vertical and horizontal axes represent canonical scores for each function. The density normal curves of F4 and F5 show that groups are not well separated, consistent with the non-significance of these functions.

Figure 4: PBL CM^+ repertoire are significantly more perturbed than CM^- PBL or spleen repertoires and can be clustered separately. (a) BV-BC perturbations (DBV-BC) were computed with ISEApeaks using the CTR spleen group as control. Mean sample DBV-BC (μDBV-BC) with their standard error are shown for CM^+, CM^- and CTR mice in the PBL and spleen. DBV-BC range from 0, identical to the reference repertoire, to 100, completely perturbed. (b) Schematic representation of sample clusters obtained with k-mean clustering on DBV-BC with k=4 and 3. BV-BC perturbation of each sample was analyzed without prior knowledge of group composition. For k=5 or 6, PBL CM^+ samples were split in two clusters.

(c) The BV-BJ repertoires of PBL samples (6 CTR, 5 CM$^-$ and 6 CM$^+$) were correctly grouped by k-mean clustering with k=3 on DBV-BJ data.

Figure 5: BV-BJ CDR3 profiles of recurrent expansions identified by OligoScore and perturbation analysis. PBL cDNA were amplified with BV8.1- or BV2-specific primers with a BC-specific primer. PCR products were then subjected to run-off with appropriate BJ-specific primers. Products were separated on an automated sequencer and analyzed with Immunoscope and ISEApeaks. All PBL samples for which those combinations were analyzed are represented. Horizontal axis represents the nucleotide size and vertical axis the fluorescence intensity, in arbitrary units.

Figure 6: ISEApeaks architecture and data flowchart. ISEApeaks modules and specific files are indicated in shaded rectangles.

Figure 7: ISEApeaks peak smoothing and quality checks. (A) A CDR3 profile of human BV4-BC amplification as provided by the Immunoscope product. Nine peaks are visible by eye. (B) Histograms of peaks as obtained after extraction with DataExtractor (raw Immunoscope peak data) and after application of the three filters of DataSmoother. After filtering, peaks correspond to what is seen on the Immunoscope profile. (C) Table of peak data of this BV4-BC profile after each filter as it appears in the DataFormatter file. DataFormatter highlights adjacent peaks in light grey, ambiguous peaks in dark grey, the main peak in medium grey and the first peak that can be attributed to the mTheoreticLength (here 192 nt) in bold. Note that ISEApeaks uses color. In this example, DataSmoother finally solved all problems: each peak can be attributed to a theoretical length without conflicts.

Figure 8: Separation of samples according to BV-BC repertoire. Discriminant Analysis was performed on the new set of variables obtained with Principal Components Analysis. Bivariate graphics represent the sample value in two discriminant functions. Discriminant functions are linear combination of the variables that endeavor to separate groups. F1 to F3 stands for each discriminant function, sorted decreasingly by the associated eigenvalue. The first two functions are statistically significant, indicating that only three groups can be separated ($p<0.01$). Each sample is represented on the bivariate plots, but are scarcely visible since DA groups the samples very well. 0.95 confidence ellipses, centered on group centroid, are overlaid. The density normal curve of each group is shown on the diagonal. Vertical and horizontal axes represent canonical scores for each function. The density normal curve of F3 shows that groups are not well separated, consistent with the non-significance of this function. Sample numbers are as follows: 8 CTR PBL, 6 CTR spleen, 10 HP PBL and 8 HP spleen.

Figure 9: Representation of BV-BC perturbations. BV-BC perturbations for each BV segments (horizontal axis) and each sample (vertical axis) have been color-coded. Mice ID are indicated in the third column. TCRB perturbations were computed and displayed with ISEApeaks product. Coding is as follows: light grey (DBV-BC<5), mid grey (<10), dark grey (<20), pink (<25), red (<30), dark red (<50), black (<100). "excl" denotes excluded combinations for which the recovered signal was too poor.

Figure 10: Perturbation of TCRB repertoires induced by *P. berghei* ANKA. (A) BV-BC repertoire perturbations (DBV-BC) during hyperparasitemia. DBV-BC were computed with ISEApeaks using the CTR spleen group as control. Mean sample DBV-BC (μDBV-BC) with their standard error are shown for HP and CTR mice in the PBL and spleen. DBV-BC

range from 0, identical to the reference repertoire, to 100, completely perturbed. Sample numbers are identical to Figure 8. (B) Schematic representation of sample clusters obtained with k-mean clustering on DBV-BC with k=2.

Figure 11: DBV-BJ perturbation during hyperparasitemia. The BV-BJ repertoires of eleven PBL samples (6 CTR and 5 HP) were correctly grouped by k-mean clustering with k=2 on DBV-BJ data.

Figure 12: CDR3 profiles of BV2-BJ1.3 and BV2 BJ1.1 in PBL of HP and CTR mice. PBL cDNA were amplified with BV2- and a BC-specific primer. PCR products were then subject to run-off with appropriate BJ-specific primers. Products were separated on an automated sequencer and analyzed with the Immunoscope and ISEApeaks packages. Horizontal axis represents the nucleotide size and vertical axis the fluorescence intensity in arbitrary units.

Figure 13: Organisation scheme of the ISEApeaks package.

Figure 14: ISEApeaks main menu.

Figure 15: ISEApeaks Preferences menu.

Figure 16: ISEApeaks menu bar.

Figure 17: A DataParameter CGEL worksheet.

Figure 18: A DataParameter CPICTPLACES worksheet.

Figure 19: A DataFormatter file.

Figure 20: DataAnalyser 'para' worksheet is used to parameterise the different macros of DataAnalyser.

Figure 21: A DataAnalyser 'Peaks' worksheet.

Figure 22: A DataAnalyser worksheet showing the result of the 'PercentImport' macro.

Figure 23: the 'Perturbation1' macro.

Figure 24: the 'Perturbation2' macro.

Figure 25: the 'RIS' macro.

Figure 26: the 'OligoScore' macro.

Figure 27: the 'DrawArray' procedure creates a representation of diversity in the repertoires.

DETAILED DESCRIPTION OF THE INVENTION

Unless specifically defined, all technical and scientific terms used herein have the same meaning as commonly understood by a skilled artisan in computers, software/program design, biochemistry, cellular biology, molecular biology, and the medical sciences.

All methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, with suitable methods and materials being described herein. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In case of conflict, the present specification, including definitions, will control. Further, the materials, methods, and examples are illustrative only and are not intended to be limiting, unless otherwise specified.

In the central paradigm of modern molecular biology, biological information flows from DNA to RNA to protein. This general scheme gives rise to a powerful template-driven precision, in which the experimenter has the ability to manipulate any one of these classes of biomolecules based on the knowledge of another. Moreover, patterns of sequence homology and relatedness are tools to predict function and reveal evolutionary relationships. With the recent completion of the genomic sequences of humans and several other commonly studied model organisms, researchers are in the unprecedented position to probe the intricate interrelationship of biomolecules, not just in the context of this paradigm, but also in the context of gene identification, protein expression, characterization of protein activity, and characterization of protein-protein interactions, protein-ligand interactions, protein-drug interactions, and protein-DNA/RNA. Accordingly, researchers have devised a dynamic array of techniques to probe the structure, activity, and relationship of DNA, RNA, and proteins.

The present invention provides, in part, a program for handling the extensive amounts raw data made available by automated sequencers and raw data extraction programs. The automated sequencers and raw data extraction programs have been designed to digitize the results of growth culturing, electrophoresis, chromatography, blotting techniques, centrifugation, DNA microarrays, or protein microarrays or sugar residue arrays of biological samples. The nature of position and area of peaks will depend on the nature of the new data produced by these techniques, either periodical or ordered. For example, in the case of microarrays, peak position will correspond to its coordinate (column x row) on the array; the peak area will correspond to the intensity of the signal detected.

In addition, the present invention provides a novel program for efficient retrieval of this raw data which has been extracted by automated sequencers and raw data extraction programs.

Moreover, the present invention provides a novel program to analyze the extensive amounts of raw data extracted by automated sequencers and raw data extraction programs.

In particular, raw data extraction programs and DNA automatic sequencers are used to determine DNA fragment lengths in a wide array of applications: DNA sequencing, microsatellites, Single Nucleotide Polymorphism, Restriction or Amplified Fragment Length Polymorphism, Single Strand Conformation Polymorphism, gene expression quantification and analyses of the immune receptor diversity. All of these applications require access to raw data (peak area and nucleotidic length). Raw data being stored in one file per lane, studies rapidly give rise to hundreds of files. However, with the increasing number of samples analyzed, no tool is currently available to allow the extensive and efficient retrieval of this raw data. The Applicants have developed the ISEApeaks software package in order to satisfy these needs.

In general, ISEApeaks extracts raw data and transfers it into one Excel file per sequencing gel. Then, data of different samples can be gathered in a peak database. Raw data extraction is currently possible with data generated from bioinformatics tools including the programs IMMUNOSCOPE^{TM} (Pannetier et al, Proc Natl Acad Sci USA 1993; 90:4319-

4323, available from INSERM, Paris, France), GENESCAN™ (PE Applied Biosystems, Foster City, CA, USA), and GENOTESTER™ (Amersham, Uppsala, Sweden), the most popular packages used to determine peak area and size. Programs which are used for raw data acquisition from growth cultures, electrophoretic samples, chromatographic columns, blotting membranes, centrifugation tubes, or microarray chips, also include IMAGEQUANT™ (Molecular Dynamics, Piscataway, NJ, USA), EAGLESIGHT™ (Stratagene, La Jolla, CA, USA), QUANTITYONE™ (BioRad, Hercules, CA, USA), and MICROARRAY SUITE (Affymetrix, Santa Clara, CA, USA).

Accordingly, within the purview of the prevent invention, the primary source of the raw data is not limiting. Exemplary primary sources include growth cultures, electrophoretic samples, chromatographic columns, blotting membranes, centrifugation tubes, or microarray chips.

The means of extraction of raw data from the primary source varies with the technique or reporting system, as well as the automated sequencer and bioinformatics tool employed. These techniques are well known in the art. The biomolecules may be labeled or unlabeled. Extraction of raw DNA and RNA data may be accomplished, for example, by fluorescence by way of a fluorophore coupled to a DNA molecule, fluorescence using a DNA intercalation agent, or autoradiography in which the DNA is end-labeled with a radioisotope. Similarly, raw protein data extraction may be accomplished, for example, by using reporters such as fluorescence labeling, autoradiography, immunography, chemiluminescence.

In one embodiment of the present invention is a method for high throughput analysis of data sets generally described by sets of peaks having a defined position and area. In this embodiment, the data may be extracted from a primary source by a bioinformatics tool (i.e. raw data extraction program or DNA automatic sequencer) and the resultant peaks are subsequently smoothed according to a user-defined parameter file. Each smooth peak data set is then stored in a data file.

In another embodiment, the particular profiles (data files), representing peaks, can be recreated for further analysis.

In a further embodiment, the peaks present in the individual data files can be assimilated into a single peak database using ISEApeaks.

In yet another embodiment, the peak database is analyzed by statistical tools contained within the ISEApeaks program.

In another embodiment of the present invention is a method for determining prognostic and diagnostic criteria by high throughput analysis of data sets generally described by sets of peaks having a defined position and area by extracting the data from a primary source by a bioinformatics tool (i.e. raw data extraction program or DNA automatic sequencer) and subsequently smoothing the resultant peaks according to a user-defined parameter file. Each smooth peak data set is then stored in a data file, which contains a particular profiles (data files), representing peaks, that is recreated for further analysis. The peaks present in the individual data files can then be assimilated into a single peak database using ISEApeaks and analyzed by statistical tools contained within this program.

The prognostic and diagnostic criteria that are established by this method can be used in field of physiopathology such as immunotheraphy, cancer treatment, HIV, infectious disease, and autoimmune disease.

The methods of the present invention may be used as a high throughput method for analysis of immune repertoires.

Immune repertoires of T or B cells are very often studied by CDR3 spectratyping. However, data obtained with this method is usually subject to a biased eye analysis. Accordingly, the ISEApeaks software package provided herein has been employed to retrieve and handle peak data from automated sequencers, from which CDR3 spectratype data is obtained. In a general strategy two new specific modules and multivariate statistics analyses are used to analyze the CDR3 spectratype. The first module addresses the crucial problem of peak smoothing. The second is a toolbox for the analysis of CDR3 spectratypes, which includes perturbation computation, recurrent peak finding, expansion assessment and datamining. To illustrate this approach, the complex TCRB repertoire modifications induced by *Plasmodium berghei* ANKA infection were assessed and are presented in Example 2. This global and exhaustive repertoire analysis approach is of general interest for T and B lymphocyte repertoire studies and is currently used in human cohorts in various pathologies and during clinical trials.

Roles of B or T cells in abnormal situations, such as infectious diseases (Louis et al, *Curr. Opin. Immunol.* 1998; 10:459 and Boubou et al, *Int. Immunol.* 1999; 11:1553), autoimmunity (Wagner et al, Proc. Natl. Acad. Sci. USA 1998; 95:14447) or cancers (Pannetier et al, The Immunoscope technique for analysis of TCR repertoire. In: The human antigen T cell receptor: selected protocols and applications. 1995 JR Oksenberg, Austin, Texas, p. 287), are widely studied by examining their repertoire of antigen-specific receptors. During lymphocyte differentiation, the diversity of these heterodimeric receptors is produced by random somatic DNA rearrangements of V, (D) and J segments later spliced to C segments (Davis and Bjorkman, *Nature* 1988; 334:395). The product of the V(D)J joining, called the Complementary Determining Region 3 (CDR3), is in contact with the antigen. This region is imprecise in the number and the nature of nucleotides that are removed or added and is therefore variable in amino-acid length and composition.

Several approaches have been used to describe a repertoire. One of the most widely used is CDR3 spectratyping that describe the diversity of a T cell population repertoire by the analysis of the CDR3 length distribution (Pannetier et al, *Proc. Natl. Acad. Sci. USA* 1993;

90:4319). Briefly, sets of PCR amplification are performed with V-specific and C- or J-specific primers. These PCR products are then labeled in run-off experiment with C- or J-specific primers coupled with a fluorophore and loaded on an automated DNA sequencer to separate the different CDR3 lengths in each V-C or V-J combination. GENESCAN™ (Applied Biosystems, Foster City), IMMUNOSCOPE™ (INSERM, Paris) and GENOTESTER™ (Amersham, Uppsala) are three popular software packages used to determine nucleotide sizes and areas of the observed CDR3 peaks. This CDR3 spectratyping technique, by far the more exhaustive one, can for instance describe the human TCRB repertoire with up to 2400 measurements (Pannetier et al, *Immunol Today* 1995; 16:176). In the past, CDR3 spectratype data has mainly been analyzed qualitatively by eye, which can introduce biases in analysis and possibly lead to loss of relevant information. With the help of recent automated sequencers (96 well- and/or capillary-based sequencers), it is now possible to analyze cohorts of individuals. Research teams performing such high throughput acquisition are rapidly overwhelmed by the amount of data. A complete use of the CDR3 spectratype method thus requires the development of appropriate software tools for retrieving, handling, organizing and objectively analyzing data scattered through dozens of sample files on different gels.

Accordingly, the present Inventors have developed the ISEAPEAKS™ software package to retrieve, handle and organize raw data generated by bioinformatics tools, such as GENESCAN™, IMMUNOSCOPE™ (Collette and Six, *Bioinformatics* 2002; 18:329) and GENOTESTER™ software. Data of different samples can be gathered in an EXCEL™ (Microsoft, Seattle, Washington, USA) peak database. According to the present invention, the EXCEL™ peak database was used to set up an original strategy based on multivariate statistics to achieve a global description of CDR3 repertoires.

To illustrate this strategy, TCRB repertoire data from *Plasmodium berghei* ANKA (PbA) infected B10.D2 mice were analyzed to characterize modifications during malaria (Example 2). In this model, PbA induces either fatal cerebral malaria between day 7 to 10 after infection or severe anemia due to hyperparasitemia (HP) leading to death three weeks after infection (Boubou et al, *Int. Immunol.* 1999; 11:1553). In Example 2, only mice presenting HP were considered. Mice developing the cerebral syndrome were not included in Example 2, but are described in Example 1.

In another object of the present invention is a high throughput method for analysis of immune repertoires by starting with biological samples, which contain DNA or RNA fragments. The DNA or RNA fragments are then purified. If the source is purified RNA, cDNA is then transcribed by using reverse transcription PCR (Sambrook et al, Molecular Cloning: A Laboratory Manual, 3rd Edition, 2001, Cold Spring Harbor Laboratory Press). The purified DNA or cDNA, obtained by transcribing the RNA, are then amplified by PCR or strand displacement amplification (SDA, Walker et al, *Clin Chem* 1996; 42:9-13 and Little et al, *Clin Chem* 1999; 45:777-784) methods using oligonucleotides specific for antigen specific receptor genes, for example Immunoglobulin and T-cell receptor, variable (V), Junctional (J) and Constant (C) regions. The amplified DNA are labeled for detection. Suitable DNA labels include radiolabels (available from ICN Biomedicals, Costa Mesa, CA, USA) chemiluminescent probes, and fluorophores (a broad range are available from Molecular Probes, Eugene, OR, USA), for example performing a runoff extension step with J or C specific oligonucleotide labeled with a fluorescent drug. Each labeled amplified DNA is then electrophoretically separated in an automatic sequencer and the eletrophoregram is analyzed, identifying peaks by determining their position and area that correspond to labeled amplified DNA.

The sets of peaks having a defined position and area may be extracted from the electrophoregram by a bioinformatics tool (i.e. raw data extraction program or DNA automatic sequencer) and subsequently smoothing the resultant peaks according to a user-defined parameter file. Each smooth peak data set is then stored in a data file, which contains a particular profiles (data files), representing peaks, that is recreated for further analysis. The peaks present in the individual data files can then be assimilated into a single peak database using ISEApeaks and analyzed by statistical tools contained within this program.

As is evident from the central paradigm of modern molecular biology, the samples that are amenable to the present invention are the fundamental biomolecules include DNA, RNA, and proteins.

The proteins embraced by the present invention encompass a vast spectrum of protein classes, all of which are within the purview of the present invention. Some examples of protein-types include, but by no means are limited to, antigenic proteins, antibodies, DNA-binding proteins, RNA-binding proteins, kinases, methylases, proteases, proteins involved in replication, proteins involved cell division, and proteins involved in regulation of cellular processes and homeostasis.

As used herein the term "biological sample" refers to a solution, mixture, or suspension that contains one or more biomolecule. The term "biomolecule" refers to any matter of biological origin, which includes intact cells, cellular material, DNA, RNA and proteins. These biomolecules may be naturally occurring or may be obtained by synthetic methods. In addition, a biomolecule obtained from either means may be modified or unmodified and may be a member of an extract or may be isolated or purified.

The term "naturally occurring" refers to a DNA, RNA, or protein that has is found in or expressed from a living host organism. This term includes genomically, chromosomally, plasmid, and cosmid expressed proteins, as well as genomic DNA, chromosomal DNA, plasmid DNA and cosmid DNA. Accordingly, this term also embraces the RNA the is transcribed from genomic DNA, chromosomal DNA, plasmid DNA and cosmid DNA.

The term "synthetic methods" in relation to DNA or RNA refers to solid-phase or liquid-phase synthesis. Examples of suitable synthetic methods for DNA or RNA are provided by Rayner et al (*Genome Res*. 1998; 8:741-747), Lashkari et al (*Proc Natl Acad Sci* 1995; 92:7912-7915), and Andrus et al (*Nucleic Acids Symp Ser* 1995; 34:183-184). In relation to proteins the term "synthetic methods" refers to *in vitro* transcription and translation of target DNA to protein in a single tube. Two common commercially available kits to for in vitro protein system are the EXPRESSWAY™ System (Invitrogen Life Technologies, Carlsbad, CA, USA) and TNT™ Quick Coupled Transcription/Translation System (Promega, Madison, WI, USA).

As used herein, the term "unmodified" relates to DNA, RNA, or protein molecules that exist in their nascent state, i.e. does not have any post-replication, post-transcription, or post-translation alterations (deletions, mutations, additions).

As used herein, the term "modified" relates to DNA, RNA, or proteins molecules, which have been altered post-replicationally, post-transcriptionally, or post-translationally. These modifications may occur within the host source or by experimental manipulation. Common modifications of DNA and RNA include methylation, acetylation, nucleoside excision, fluorophore labeling, and radiolabeling. Similarly, common protein modifications include methylation, acetylation, nucleoside excision, fluorophore labeling, radiolabeling, carbohydrates, glycoconjugates, and lipids. A detailed description and understanding can be readily obtained for carbohydrate, glycoconjugate, and lipid protein modification can be obtained by reference to, *inter alia*, Essentials of Glycobiology (1999), Edited By Ajit Varki, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. Suitable fluorophores for DNA, RNA, and proteins, and detailed descriptions thereof, are available in the

Handbook of Fluorescent Probes and Research Products from Molecular Probes (Eugene, OR, USA).

Within the context of the present invention "isolated" or "purified" means separated out of its natural environment, which is also substantially free of other contaminating proteins, polynucleotides, and/or other biological materials often found in cell extracts. Techniques for obtaining cell extracts are described in Roe et al (DNA isolation and Sequencing, 1996, John Wiley & Sons, New York), Sambrook et al (Molecular Cloning: A Laboratory Manual, 3rd Edition, 2001, Cold Spring Harbor Laboratory Press), and Scopes (Protein Purification, Principles and Practice, 1994, Springer-Verlag New York, Inc.). Common methods of isolation and/or purification of DNA, RNA, and/or proteins include: centrifugation, precipitation, batch adsorption, chromatography, and electrophoresis (Scopes, Protein Purification, Principles and Practice, 1994, Springer-Verlag New York, Inc.).

Chromatographic techniques that are routinely used to isolate and/or purify DNA, RNA, and/or proteins include, liquid chromatography and column chromatography.

"Liquid Chromatography" in the context of this invention includes, High-Performance Liquid Chromatography (HPLC), Reverse Phase-HPLC, Fast Performance Liquid Chromatography (FPLC). The artisan is directed to Scopes (Protein Purification, Principles and Practice (1994), Springer-Verlag New York, Inc.) for a more detailed description of these techniques.

"Column Chromatography" in the context of this invention includes, ion-exchange chromatography, inorganic adsorbents, hydrophobic adsorbents, immobilized metal affinity chromatography (IMAC), cationic polymer-nucleic acid complexes, thiophilic adsorbents, mixed-function adsorbents, affinity chromatography, immunoadsorbent chromatography, dye-ligand chromatography, and gel filtration chromatography. The artisan is directed to

Scopes (Protein Purification, Principles and Practice (1994), Springer-Verlag New York, Inc.) for a more detailed description of these techniques.

Electrophoretic techniques that routinely used to isolate and/or purify DNA, RNA, and/or proteins include, native gel electrophoresis, urea gel electrophoresis, sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), gradient gel electrophoresis, isoelectric focusing, two-dimensional gel electrophoresis, and capillary electrophoresis. Additional electrophoretic techniques may include coupled gel-membrane techniques for isolating, separating, and/or probing DNA, RNA, or proteins, such as Southern Blotting, Northern Blotting, and Western Blotting. The artisan is directed to Scopes (Protein Purification, Principles and Practice, 1994, Springer-Verlag New York, Inc.) and Ausubel et al (Current Protocols in Molecular Biology, 1994, New York: Greene Publishing Assoc. and Wiley-Interscience) for a more detailed description of these techniques.

The term "growth culture" as used herein refers to bacterial, phage, viral, and eukaryotic growth in solid, liquid, or gaseous medium.

Figure 1 illustrates a computer system 1101 upon which an embodiment of the present invention may be implemented. The computer system 1101 includes a bus 1102 or other communication mechanism for communicating information, and a processor 1103 coupled with the bus 1102 for processing the information. The computer system 1101 also includes a main memory 1104, such as a random access memory (RAM) or other dynamic storage device (e.g., dynamic RAM (DRAM), static RAM (SRAM), and synchronous DRAM (SDRAM)), coupled to the bus 1102 for storing information and instructions to be executed by processor 1103. In addition, the main memory 1104 may be used for storing temporary variables or other intermediate information during the execution of instructions by the processor 1103. The computer system 1101 further includes a read only memory (ROM) 1105 or other static storage device (e.g., programmable ROM (PROM), erasable PROM

(EPROM), and electrically erasable PROM (EEPROM)) coupled to the bus 1102 for storing static information and instructions for the processor 1103.

The computer system 1101 also includes a disk controller 1106 coupled to the bus 1102 to control one or more storage devices for storing information and instructions, such as a magnetic hard disk 1107, and a removable media drive 1108 (e.g., floppy disk drive, read-only compact disc drive, read/write compact disc drive, compact disc jukebox, tape drive, and removable magneto-optical drive). The storage devices may be added to the computer system 1101 using an appropriate device interface (e.g., small computer system interface (SCSI), integrated device electronics (IDE), enhanced-IDE (E-IDE), direct memory access (DMA), or ultra-DMA).

The computer system 1101 may also include special purpose logic devices (e.g., application specific integrated circuits (ASICs)) or configurable logic devices (e.g., simple programmable logic devices (SPLDs), complex programmable logic devices (CPLDs), and field programmable gate arrays (FPGAs)).

The computer system 1101 may also include a display controller 1109 coupled to the bus 1102 to control a display 1110, such as a cathode ray tube (CRT), for displaying information to a computer user. The computer system includes input devices, such as a keyboard 1111 and a pointing device 1112, for interacting with a computer user and providing information to the processor 1103. The pointing device 1112, for example, may be a mouse, a trackball, or a pointing stick for communicating direction information and command selections to the processor 1103 and for controlling cursor movement on the display 1110. In addition, a printer may provide printed listings of data stored and/or generated by the computer system 1101.

The computer system 1101 performs a portion or all of the processing steps of the invention in response to the processor 1103 executing one or more sequences of one or more instructions contained in a memory, such as the main memory 1104. Such instructions may be read into the main memory 1104 from another computer readable medium, such as a hard disk 1107 or a removable media drive 1108. One or more processors in a multi-processing arrangement may also be employed to execute the sequences of instructions contained in main memory 1104. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions. Thus, embodiments are not limited to any specific combination of hardware circuitry and software.

As stated above, the computer system 1101 includes at least one computer readable medium or memory for holding instructions programmed according to the teachings of the invention and for containing data structures, tables, records, or other data described herein. Examples of computer readable media are compact discs, hard disks, floppy disks, tape, magneto-optical disks, PROMs (EPROM, EEPROM, flash EPROM), DRAM, SRAM, SDRAM, or any other magnetic medium, compact discs (e.g., CD-ROM), or any other optical medium, punch cards, paper tape, or other physical medium with patterns of holes, a carrier wave (described below), or any other medium from which a computer can read.

Stored on any one or on a combination of computer readable media, the present invention includes software for controlling the computer system 1101, for driving a device or devices for implementing the invention, and for enabling the computer system 1101 to interact with a human user. Such software may include, but is not limited to, device drivers, operating systems, development tools, and applications software. Such computer readable media further includes the computer program product of the present invention for performing all or a portion (if processing is distributed) of the processing performed in implementing the invention.

The computer code devices of the present invention may be any interpretable or executable code mechanism, including but not limited to scripts, interpretable programs, dynamic link libraries (DLLs), Java classes, and complete executable programs. Moreover, parts of the processing of the present invention may be distributed for better performance, reliability, and/or cost.

The term "computer readable medium" as used herein refers to any medium that participates in providing instructions to the processor 1103 for execution. A computer readable medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media includes, for example, optical, magnetic disks, and magneto-optical disks, such as the hard disk 1107 or the removable media drive 1108. Volatile media includes dynamic memory, such as the main memory 1104. Transmission media includes coaxial cables, copper wire and fiber optics, including the wires that make up the bus 1102. Transmission media also may also take the form of acoustic or light waves, such as those generated during radio wave and infrared data communications.

Various forms of computer readable media may be involved in carrying out one or more sequences of one or more instructions to processor 1103 for execution. For example, the instructions may initially be carried on a magnetic disk of a remote computer. The remote computer can load the instructions for implementing all or a portion of the present invention remotely into a dynamic memory and send the instructions over a telephone line using a modem. A modem local to the computer system 1101 may receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to the bus 1102 can receive the data carried in the infrared signal and place the data on the bus 1102. The bus 1102 carries the data to the main memory 1104, from which the processor 1103 retrieves and executes the instructions. The instructions received by the main memory 1104 may optionally be stored on storage device 1107 or 1108 either before or after execution by processor 1103.

The computer system 1101 also includes a communication interface 1113 coupled to the bus 1102. The communication interface 1113 provides a two-way data communication coupling to a network link 1114 that is connected to, for example, a local area network (LAN) 1115, or to another communications network 1116 such as the Internet. For example, the communication interface 1113 may be a network interface card to attach to any packet switched LAN. As another example, the communication interface 1113 may be an asymmetrical digital subscriber line (ADSL) card, an integrated services digital network (ISDN) card or a modem to provide a data communication connection to a corresponding type of communications line. Wireless links may also be implemented. In any such implementation, the communication interface 1113 sends and receives electrical, electromagnetic or optical signals that carry digital data streams representing various types of information.

The network link 1114 typically provides data communication through one or more networks to other data devices. For example, the network link 1114 may provide a connection to a another computer through a local network 1115 (e.g., a LAN) or through equipment operated by a service provider, which provides communication services through a communications network 1116. In preferred embodiments, the local network 1114 and the communications network 1116 preferably use electrical, electromagnetic, or optical signals that carry digital data streams. The signals through the various networks and the signals on the network link 1114 and through the communication interface 1113, which carry the digital data to and from the computer system 1101, are exemplary forms of carrier waves transporting the information. The computer system 1101 can transmit and receive data, including program code, through the network(s) 1115 and 1116, the network link 1114 and the communication interface 1113. Moreover, the network link 1114 may provide a connection through a LAN 1115 to a mobile device 1117 such as a personal digital assistant

(PDA) laptop computer, or cellular telephone. The LAN communications network 1115 and the communications network 1116 both use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on the network link 1114 and through the communication interface 1113, which carry the digital data to and from the system 1101, are exemplary forms of carrier waves transporting the information. The processor system 1101 can transmit notifications and receive data, including program code, through the network(s), the network link 1114 and the communication interface 1113.

Having generally described this invention, a further understanding can be obtained by reference to certain specific examples, which are provided herein for purposes of illustration only, and are not intended to be limiting unless otherwise specified.

EXAMPLES

Example 1.1: Mice and Parasites

Eight-week old B10.D2 mice were purchased from Harlan UK Limited. The clone 1.49L of *Plasmodium berghei* ANKA (Amani, V., M.I. Boubou, S. Pied, M. Marussig, D. Walliker, D. Mazier, and L. Renia. 1998. Cloned lines of Plasmodium berghei ANKA differ in their abilities to induce Experimental Cerebral Malaria. Infect. Immun 66:4093-4099.) was kindly given by Dr. Walliker (Institute of Genetics, Edinburg, UK) and is maintained in the laboratory on C57BL/6J female mice. This clone induced in mice a neurological syndrome partly mimicking the one of human CM. Erythrocytic stages of the parasite were cryopreserved in liquid nitrogen as stabilates in Alserver's solution containing 10% glycerol. Infection was induced by intraperitoneal injection of 10^6 parasitized red blood cells. Between day 7 to day 10 after infection, 90% of the mice developed cerebral malaria characterized by ataxia, paralysis, deviation of the head and convulsions followed by deep coma and death. These mice constituted the CM$^+$ group. CM$^-$ mice were sacrificed between day 11 to 16 after infection because they did not shown signs of CM during the critical period, however, they did exhibit a parasitemia above 20%.

Example 1.2: Cell preparation

Blood was obtained on heparin by retroorbital punction. Mononuclear cells were isolated on ficoll-Hypaque gradient (Pharmacia, France). Spleen was removed and cells suspended in 3% FCS-PBS. Red blood cells were lysed with ammonium chloride buffer (ACK) for five minutes at room temperature. Cell preparations were then washed twice with PBS. Lymphoid cells were counted using Malassez cell in presence of eosin to exclude dead cells.

Example 1.3: TCRB repertoire

Total RNA was extracted from more than 90,000 mononuclear cells for each sample using the TRI REAGENT kit (Molecular Research Center, Cincinnati, Ohio). 20 µg of glycogen (Roche, Meylan, France) was used to ensure optimal precipitation of RNA and pellet visualization. Protocols for TCR BV-BC and BV-BJ CDR3 spectratyping have been described previously, which were utilized in the present example (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.; Pannetier, C., J. Even, and P. Kourilsky. 1995. The Immunoscope technique for analysis of TCR repertoire. In The human antigen T cell receptor: Selected protocols and applications. J. R. Oksenberg, Austin, Texas. 287-325.). BC, BV and BJ primer sequences were those described previously

(Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.), except BV8.3 (5'-TGCTGGCAACCTTCAAATAGGA-3') (SEQ ID NO: 1) and BV13 (5'-AGGCCTAAAGGAACTAACTCCAC-3') (SEQ ID NO: 2). Because BV5.3 (Chou, H.S., S.J. Anderson, M.C. Louie, S.A. Godambe, M.R. Pozzi, M.A. Behlke, K. Huppi, and D.Y. Loh. 1987. Tandem linkage and unusual RNA splicing of the T-cell receptor beta-chain variable-region genes. Proc. Natl. Acad. Sci. USA 84:1992-1996.), BV17 (Wade, T., J. Bill, P.C. Marrack, E. Palmer, and J.W. Kappler. 1988. Molecular basis for the nonexpression of Vb 17 in some strains of mice. J. Immunol. 141:2165-2167.), and BV19 (Louie, M.C., C.A. Nelson, and D.Y. Loh. 1989. Identification and characterization of new murine T cell receptor b chain variable region (Vb) genes. J. Exp. Med. 170:1987-1998.) are not functional in B10.D2 mice, they were not amplified. For BV-BJ repertoires, BV2, BV3, BV4, BV5.1, BV6, BV7, BV8.1-3, BV9, BV14 and BV16 genes were analyzed. PCR products were loaded on a 36-well ABI373 automated sequencer (Applied Biosystems, Foster city, CA) and separated according to their nucleotide length forming a profile of peaks for each primer combination, spaced by 3 nucleotides as expected for in-frame transcripts. Each peak corresponded to a CDR3 length. The Immunoscope product (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.) was used to obtain peak area and nucleotide length and CDR3 profile displays from sequencer raw data.

Example 1.4: BV-BJ direct sequencing

Direct sequencing was performed with BV and BJ primers for peaks representing between 65% and 100% of the BV-BJ profile following the recommendations of the ABI Prism Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems). BV-BJ PCR products were first reamplified. PCR products were then incubated with 0.5 units of Shrimp Alkaline Phosphatase (USB) and 5 units of Exonuclease I (USB) at 37°C for 40 min followed by 20 min at 80°C. DNA alignments were performed using the GCG product.

Example 1.5: Methods and tools for CDR3 spectratype analyses

The present inventors developed the ISEApeaks® product (©2000-2002 Institut Pasteur, France, Paris) to extract, smooth, manage and analyze the large amount of data obtained in this study (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330. and Example 2). As the intensity of CDR3 peaks is not comparable between different amplifications, the percentage of use of each CDR3 length was obtained by dividing the area of CDR3 peaks by the total area of all peaks within the profile.

Perturbation of BV-BC repertoire were compared to a control group as explained previously (Gorochov, G., A.U. Neumann, A. Kereveur, C. Parizot, T.S. Li, C. Katlama, M. Karmochkine, G. Raguin, B. Autran, and P. Debre. 1998. Perturbation of CD4+ and CD8+ T-Cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. Nat. Med 4:215-221; Han, M., L. Harrison, P. Kehn, K. Stevenson, J. Currier, and M.A. Robinson. 1999. Invariant or highly conserved TCRα are expressed on double-negative (CD3+CD4-CD8-) and CD8+ T cells. J. Immunol. 163:301-311.). Sample perturbation (μDBV-BC) is the mean of perturbations of each BV segment (DBV-BC). The perturbation index was expanded to BV-BJ repertoires by computing μμDBV-BJ, the mean of the DBV-BJ for all BV and BJ segments. BV-BC and BV-BJ perturbations range from 0, identical profiles, to 100, completely different profiles. For sake of reference, TCRa, TCRb, TCRg, and TCRd refer to TCRα, TCRβ, TCRγ, and TCRΔ, respectively. In addition TCRab refers to TCRαβ and TCRgd refers to TCRγΔ.

Recurrence of CDR3 expansions was assessed quantitatively with OligoScore (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.), which scored each peak in each group of samples. The maximum score of the control group was used as a threshold for other groups. Peaks with a score above this threshold were considered recurrently expanded.

Example 1.6: Statistics

Multivariate statistics were used to analyze repertoire data. In a first approach, each repertoire was considered as a vector in n-dimensional space, where n is the number of variables that describe the repertoires. Missing values were replaced by the overall mean of the variable (Rencher, A.C. 1995. Methods of multivariate analysis. J Wiley, New York. 627 pp.). The number of variables (230 peaks for BV-BC repertoires) was too high for theoretical constraints of Discriminant Analysis and was thus reduced using Principal Components Analysis (PCA). PCA extracts new variables from the data set that retains the variability contained in the original data set. Linear Discriminant Analysis (DA) was performed on the new data set to compare the different groups. DA computes discriminant functions that maximize inter-group variation and minimize intra-group variation. Significance of each discriminant function was tested using χ^2 approximation of Wilks' statistics (Rencher, A.C. 1995. Methods of multivariate analysis. J Wiley, New York. 627 pp.).

Univariate one-way or two-way Analysis of Variance (ANOVA) was used to analyze sample perturbation data (μDBV-BC or μμDBV-BJ). When significant, comparison between two categories was performed with Fischer's Protected Least Significant Difference. One-way Multiple Analysis of Variance (MANOVA) was used to compare DBV-BJ. MANOVA F-statistics approximation of Wilk's lambda, Roy's greatest root, Pillai trace and Hotelling-Lawley trace multivariate statistics were calculated. As the power of these four statistics are not equivalent (Rencher, A.C. 1995. Methods of multivariate analysis. J Wiley, New York. 627 pp.), MANOVA was considered significant when all four statistics were significant. Indicated p value corresponds to the maximum value of the p value obtained for these four statistics (Spanakis, E., and D. Brouty-Boye. 1997. Discrimination of fibroblast subtypes by multivariate analysis of gene expression. Int. J. Cancer 71:402-409.). When significant, a one-way ANOVA was carried out on each variable (BV-BC perturbation) to assess for differences in the considered groups. This enables minimization of global error rate of the univariate tests (Rencher, A.C. 1995. Methods of multivariate analysis. J Wiley, New York. 627 pp.).

To further assess similarity between samples, k-mean clustering with Euclidean distance was used. Significance of the clusters were assessed by computing the probability of observing by chance the number of samples of a particular group within the cluster (Tavazoie, S., J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. 1999. Systematic determination of genetic network architecture. Nat. Genet 22:281-285.).

PCA, DA and k-mean clustering were performed with SYSTAT 10 (SPSS). ANOVA and MANOVA were performed with StatView 5.0 (SAS Institute Inc.). Statistics were considered significant when $p<0.01$.

Example 1.7: Analysis of TCRB repertoire during Cerebral Malaria by the Immunoscope Method.

The TCRB repertoire during Cerebral Malaria (CM) induced in B10.D2 mice by *Plasmodium berghei* ANKA was analyzed using the exhaustive CDR3 length spectratyping approach (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.).BV-BC repertoires were studied with total cDNA, prepared from non-infected control (CTR) mice and infected mice that did (CM^+) or did not (CM^-) develop CM. In PBL and spleen of naive mice, CDR3 profile for each BV-BC combination was bell-shaped, indicative of a diverse polyclonal repertoire (Figure 2 and data not shown). The TCRB repertoires of the CM^+ and CM^- mice were profoundly altered: almost all CDR3 profiles were different from a bell-shaped profile and multiple expansions were evidenced (Figure 2). Among these complex repertoire modifications, some could contribute to the protective response against *P. berghei* ANKA infection while others may be involved in pathology. The comparison between CDR3 profiles from CM^+ and CM^- mice enabled the identification of potentially pathogenic clones associated with CM that would be present in CM^+ mice but absent in CM^- mice.

Example 1.8: Specific alterations of TCRB repertoire during cerebral malaria.

To identify CM-associated alterations, *ex vivo* PBL and splenocyte TCRB repertoires of several individuals for CTR, CM^- and CM^+ groups were analyzed. A total of 1150 BV-BC CDR3 profiles was obtained for 55 samples. The analysis of this set of BV-BC profiles each including 6-8 peaks required the use of an original approach that combined bioinformatic tools and multivariate statistics. All peak data of TCRB repertoire was extracted, formatted and edited with the ISEApeaks product to construct the peak database. Principal Components Analysis was used to reduce the initial number of variables. The new data set retained 97% of the original information. Linear Discriminant Analysis (DA) was used to evidence global modifications due to infection in CM^+ and CM^- mice. DA determined if multivariate observations of different groups were samples of the same statistical population. All members of a given group were closely clustered together (Figure 3). Moreover, sample groups were separated in four statistically significant clusters, as only the first three discriminant functions were significant: CM^+ PBL, CM^- PBL, CM^- spleen and a group containing CM^+ spleen, CTR PBL and CTR spleen.

To further characterize the repertoires, another previously-known method was used which computed CDR3 spectratype perturbation index for each sample (Gorochov, G., A.U. Neumann, A.. Kereveur, C. Parizot, T.S. Li, C. Katlama, M. Karmochkine, G. Raguin, B. Autran, and P. Debre. 1998. Perturbation of CD4+ and CD8+ T-Cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. Nat. Med 4:215-221.). All repertoires were compared to the spleen CTR repertoires to obtain a perturbation index for each BV-BC combination (DBV-BC). µDBV-BC, the average of BV-BC perturbations, were compared by two-way Analysis of Variance (ANOVA) between compartments (PBL and spleen) and the different groups of mice (CTR, CM^- and CM^+) (Figure 4a). ANOVA compared the effect of qualitative factors on a quantitative dependent variable. Infection by *P. berghei* ANKA led to significant alteration of the TCRB perturbation both due to the groups of mice ($p<0.0001$) and to the lymphoid compartment ($p=0.0002$). PBL repertoires of CM^+ mice (mean=20.2) were more perturbed than in CM^- mice (mean=15.2; t-test, df=21, $p<0.0001$). To strengthen this observation, the k-mean clustering was used to see if groups could be separated on the basis of BV-BC perturbations without knowledge of group composition, which was the case of DA and ANOVA. For k=3 (or k=4), all 13 CM$^+$ PBL samples clustered together in a group containing 17 samples (or 16) which had the probability p=1.64E-9 (p=3.86E-10) to occur by (Figure 4b). Since PBL BV-BC repertoires appeared to be specifically altered in CM$^+$ as compared to CM$^-$ mice, further analyses were performed on PBL only.

Example 1.9: BV-BC perturbation of PBL repertoires allows prediction of CM.

An investigation to see~~ if the alteration~~ of PBL BV-BC repertoire could be used to classify samples was performed. Since the sample number is not large enough to divide it in training set and testing sets, the jackknife method was used where: each sample is left out in turn and DA is performed on the remaining samples to obtain classification functions. These functions were then used to determine to which group the left-out sample belongs.

Table 1 shows that 85% of CM$^+$ PBL samples were correctly classified, indicating that perturbation of PBL repertoire can be used to discriminate between CM$^+$ and CM$^-$ mice.

Table 1: Perturbation of BV-BC repertoires allows prediction of CM[A]

		n	CM^+		CM^-		CTR		% correct
			PBL	Spleen	PBL	Spleen	PBL	Spleen	Prediction[B]
CM^+	PBL	13	11	1	0	1	0	0	85
	Spleen	10	2	4	1	1	2	0	40
CM^-	PBL	10	1	0	5	3	0	1	50
	Spleen	8	1	2	0	5	0	0	63
CTR	PBL	8	1	1	0	0	5	1	63
	Spleen	6	0	2	1	0	0	3	50
Total			16	10	7	10	7	5	60

[A] BV-BC perturbation data was analyzed by Discriminant Analysis with the jackknife method. Each sample had been left-out in turn while classification functions were computed on the remaining samples. These functions were used to predict the class membership for this sample. Left-out cases are in row and attributed categories in columns.

[B] The percentage of correct prediction was obtained by dividing the correctly classified samples by the total number of samples.

Example 1.10: BV-BJ PBL repertoire analysis during infection.

To describe more precisely the TCRB repertoire during CM, BV-BJ repertoire analysis for 9 BV genes out of 21, which represented more than two thirds of the total TCRB repertoire, was performed. A total number of 2300 BV-BJ profiles were generated. To compare perturbation results obtained from BV-BC and BV-BJ repertoire analysis, the average DBV-BJ for all BV and BJ genes for each sample (μμDBV-BJ) was first computed.

The three PBL groups (CTR, CM^- and CM^+) were significantly different (ANOVA, p<0.0001). Again, CM^+ PBL repertoires (mean=32.3) were more altered than those of CM^- PBL (mean=24.5; p=0.0035). By using k-mean clustering with DBV-BJ data, all three groups could also be reconstituted without prior knowledge of group composition (Figure 4c). 5 out of 6 CM^+ PBL samples (p=0.0001) clustered together apart from CTR and CM^- samples.

To determine which BV gene(s) contributed to this difference, the perturbation of BV genes on the basis of BV-BJ repertoires (µDBV-BJ perturbation) was computed. MANOVA analysis showed that the three groups are statistically different (p=0.01). However, only µBV8.1-BJ (ANOVA, p=0.01) was significantly more altered in CM^+ mice compared to CM^- mice. Finally, analysis of the contribution of BJ segments in BV8.1-positive T cells by MANOVA followed by ANOVA implicated only DBV8.1-BJ1.1, DBV8.1-BJ1.6, DBV8.1-BJ2.1 and DBV8.1-BJ2.2.

ISEApeaks was used to compute BJ percentages for each BV gene in PBL samples. These percentages were compared by MANOVA. BJ use was not significantly different between the CTR, CM^- and CM^+ groups (data not shown). Thus, perturbations of BV-BJ CDR3 spectratypes are not correlated with modifications of the BJ use.

Example 1-11: Identification of pathogenic recurrent clones.

The PBL BV-BC and BV-BJ repertoire data was searched for pathogenic T cell clones associated with CM. Focus was given to the identification of pathogenic clones that are recurrently present in CM^+ but not in CM^- mice. OligoScore was then used, which scores peaks for their recurrence in each group (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.). For BV-BC repertoires, no peak in CM^- or CM^+ groups has a score higher than the threshold of CTR groups (data not shown). Hence, no BV-BC peak is found recurrently expanded. The same scoring approach was used to identify recurrent clones in BV-BJ repertoires. 122 peaks in the PBL CM$^+$ mice have a score above the corresponding threshold defined with the CTR PBL group. They were compared to those of the CM$^-$ group by subtracting the corresponding peak scores. Peaks were sorted decreasingly to identify the most recurrent in CM$^+$ but absent in CM$^-$. The three most differentially expressed recurrent peaks belong to BV8.1-BJ1.5 and BV8.1-BJ2.2 profiles (Table 2 and Figure 5). BV2-BJ1.3 peak is given as an example of a peak that was recurrently expanded both in CM$^+$ and CM$^-$ mice and thus low ranked in Table 2 (this point will be discussed later). BV8.1 peaks were directly sequenced. For each of the two BV8.1-BJ combinations, similar amino-acid sequences were found in several CM$^+$ individuals (Table 3). On the contrary, direct sequencing for the PCR products in CM$^-$ samples yielded no readable CDR3 sequence (data not shown).

Table 2: Identification of recurrent expansions in CM$^+$ mice in BV-BJ CDR3 profiles by OligoScore[A]

Rank	BV	BJ	CDR3 (aa)	OS CM$^+$	OS CM$^-$	ΔCM+/-
1	8.1	2.2	10	10.25	0.25	10.01
2	8.1	1.5	9	7.51	0.44	7.07
3	8.1	2.2	9	7.03	0.18	6.85
4	2	1.1	9	7.07	1.37	5.70
5	8.1	1.3	9	5.86	0.30	5.56
6	7	1.6	9	6.07	1.58	4.49[B]
7	8.1	1.4	9	3.27	0.13	3.14
8	9	1.5	9	3.54	0.53	3.01[B]
9	8.3	2.2	10	3.30	0.30	3.00
10	6	1.5	9	3.90	0.91	2.99
59	2	1.3	9	7.01	6.12	0.53

[A] Differences (ΔCM+/-) between the OligoScore of the CM$^+$ PBL (OS CM$^+$) and the CM$^-$ (OS CM$^-$) groups were sorted decreasingly to identify recurrent CDR3 peaks differently expressed in CM$^+$ PBL but not in CM$^-$ PBL. All OligoScores can be obtained on the supplemental data web page.

[B] Only one or two CM$^-$ PBL samples were analyzable for these peaks whereas four or more were for the other peaks and groups.

Table 3: CDR3 sequences of BV-BJ expansions in CM^{+} mice.

BV	BJ	CDR3 Length (aa)	Mice		Aminoacid sequences		
					BV	<- CDR3^{A} ->	BJ
8.1	2.2	9/10	CM+	#1	CAS	SGGDXXGQL	YFG
				#2	CAS	SVGGVNTGQL	YFG
				#3	CAS	SVGQENTGQL	YFG
8.1	1.5	9	CM+	#1	CAS	SEXXDNQAP	LFG
				#2	CAS	SDGQEDQAP	LFG
				#3	CAS	SPGQDNQAP	LFG
2	1.3	9	CM+	#1	CTC	SETGSGNTL	YFG
				#7	CTC	SVTDSGNTL	YFG
				#9	CTC	SETGSGNTL	YFG
			CM-	#1	CTC	SXTGSGNTL	YFG

BV-BJ PCR products were directly sequenced for indicated PBL samples using both a BV- and a BJ-specific primers.

^{A}X stands for a position that could not be determined. Following Kabat *et al* (31), the CDR3 region was taken as encompassing amino-acids 95 to 106.

Example 1-12: Discussion

The aim of the above study was to exhaustively characterize the TCRB repertoire during cerebral malaria (CM) in B10.D2 mice infected with *Plasmodium berghei* ANKA clone 1.4. To this end, new global methods and tools were devised, based on a large-scale use of the CDR3 spectatyping approach. The results showed that the TCRB repertoire is specifically altered in the PBL of CM$^+$ mice as compared to PBL of CM$^-$ mice whereas no difference was evidenced in the spleen. This perturbation of the TCRB repertoire was partly explained by recurrent clones that are present in CM$^+$ and absent in CM$^-$ mice.

Analysis of the repertoire using the CDR3 spectratyping described the entire *ex vivo* TCRB repertoire of a sample by up to 2400 measurements. As a high number of parameters were measured, it was ensured that the cell quantity used is sufficient. In particular, paucity of material tends to favor stochastic PCR amplifications. Therefore, a set of cell dilutions were carefully determined by repertoire analysis so that the cell numbers used were sufficient to guarantee the quality of the repertoire data (A. Six *et al*, unpublished data). Furthermore, the identification of recurrent BV-BJ CDR3 expansions shows that the repertoire modifications documented herein are not artifacts (Table 3).

The original bioinformatic tools devised enabled to analyze the 3450 CDR3 profiles generated in this study. All three independent multivariate statistics were consistently gather in a similar manner the 6 experimental groups into 3 to 4 clusters. Moreover, as expected, the CTR PBL and CTR spleen groups were not separated for DA, ANOVA and k-mean clustering (with k=3). For the first time, it has been demonstrated that T cell repertoire data can give diagnostic/prognostic information when analyzed by class prediction. The alteration of the BV-BC repertoire enabled the group classification of 85% of the PBL samples of CM$^+$ mice. PBL samples of CM$^-$ mice are less correctly classified (50%). However, only one out of

10 CM$^-$ PBL samples was erroneously classified as a CM$^+$ PBL indicating that the risk to predict falsely that an infected individual is developing CM is small.

An original quantitative scoring method, OligoScore (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.), was used to identify recurrent expansion of T cell clones among the 1040 BV-BJ CDR3 peaks. It should be noted that existence of perturbation in a particular BV-BC or BV-BJ profile did not imply existence of recurrent expansion within this profile since private expansions can also distort it. Surprisingly, no recurrent peaks were found at the level of BV-BC. Two explanations can be given. Recurrent peaks in the BV-BC repertoires might have been below the detection limit of the scoring method. This is unlikely since OligoScore enabled detection of recurrence that were not visible by eye, even with a small number of samples (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330). More likely, it might have been the consequence of a "buffer effect" between BV-BC and BV-BJ data (Boudinot, P., S. Boubekeur, and A. Benmansour. 2001. Rhabdovirus infection induces public and private T cell responses in teleost fish. J. Immunol. 167:6202-6209.). Variation at the more precise level of BV-BJ repertoires could have been averaged in the corresponding BV-BC repertoires since these repertoires were the addition of all BV-BJ repertoires. An example of this "buffer effect" can be visualized on Figure 5 for BV8.1$^+$ cells where modifications seen at the BV8.1–BJ1.5 and BV8.1–BJ2.2 levels were smoothed at the BV8.1-BC level. This effect was also seen on PBL sample perturbation of CM$^+$ mice, which was 20.2 when estimated on BV-BC repertoires and 32.3 on BV-BJ repertoires. It is because BV-BJ repertoires were twelve times more precise than BV-BC and therefore gave a more accurate description of the repertoire. Finally, as BV-BJ repertoires are quantitative inside a given BV gene (Musette, P., A. Galelli, P. Truffa-Bachi, W. Peumans, P.

Kourilsky, and G. Gachelin. 1996. The Jb segment of the T cell receptor contributes to the Vb-specific T cell expansion caused by staphylococcal enterotoxin B and Urtica dioica superantigens. Eur. J. Immunol 26:618-622.; Pannetier, C., S. Delassus, S. Darche, C. Saucier, and P. Kourilsky. 1993. Quantitative titration of nucleic acids by enzymatic amplification reactions run to saturation. Nucleic Acids Res 21:577-583.), it was possible to test this "buffer effect" by calculus. BV-BC profiles were, indeed, constructed with the BV-BJ data (data not shown).

The implication of Tαβ cells in the neuropathogenesis of malaria has been demonstrated with depletion by antibodies (Curfs, J.H., T.P. Schetters, C.C. Hermsen, C.R. Jerusalem, A.A. van Zon, and W.M. Eling. 1989. Immunological aspects of cerebral lesions in murine malaria. Clin. Exp. Immunol 75:136-140., Grau, G.E., P.F. Piguet, H.D. Engers, J.A. Louis, P. Vassalli, and P.H. Lambert. 1986. L3T4+ T lymphocytes play a major role in the pathogenesis of murine cerebral malaria. J. Immunol. 137:2348-2354; Hermsen, C., T. Vandewiel, E. Mommers, R. Sauerwein, and W. Eling. 1997. Depletion Of CD4+ or CD8+ T-cells prevents Plasmodium berghei induced cerebral malaria in end-stage disease. Parasitology 114:7-12; Yanez, D.M., J. Batchelder, H.C. van der Heyde, D.D. Manning, and W.P. Weidanz. 1999. gd T-cell function in pathogenesis of cerebral malaria in mice infected with Plasmodium berghei ANKA. Infect. Immun 67:446-448.) and use of nude (Finley, R.W., L.J. Mackey, and P.H. Lambert. 1982. Virulent P. berghei malaria: prolonged survival and decreased cerebral pathology in cell-dependent nude mice. J. Immunol. 129:2213-2218.) or knockout mice (Yanez, D.M., J. Batchelder, H.C. van der Heyde, D.D. Manning, and W.P. Weidanz. 1999. gd T-cell function in pathogenesis of cerebral malaria in mice infected with Plasmodium berghei ANKA. Infect. Immun 67:446-448; Boubou, M.I., A. Collette, D. Voegtlé, D. Mazier, P.-A. Cazenave, and S. Pied. 1999. T cell response in malaria pathogenesis: selective increase in T cells carrying the TCR Vb8 during experimental cerebral malaria. Int. Immunol 11:1553-1562). The results added to these previous studies and demonstrated, *ex vivo*, a profound perturbation and the existence of recurrent CDR3 peaks in TCRB PBL repertoires during CM despite the numerous *P. berghei* molecules that stimulated the immune system during infection. By contrast with recurrent responses against single antigens (Levraud, J.P., C. Pannetier, P. Langlade-Demoyen, V. Brichard, and P. Kourilsky. 1996. Recurrent T cell receptor rearrangements in the cytotoxic T lymphocyte response in vivo against the p815 murine tumor. J. Exp. Med. 183:439-449; Faure, M., S. Calbo, J. Kanellopoulos, A.M. Drapier, P.A. Cazenave, and D. Rueff-Juy. 1999. Tolerance to maternal immunoglobulins: resilience of the specific T cell repertoire in spite of long-lasting perturbations. J. Immunol. 163:6511-6519.), the recurrent response against *P. berghei* was modest since it was not found at the BV-BC level. This could have been related to a general activation of T cells, possibly due to *Plasmodium* mitogens (Ballet, J.J., P. Druilhe, M.A. Querleux, C. Schmitt, and M. Agrapart. 1981. Parasite-derived mitogenic activity for human T cells in Plasmodium falciparum continuous cultures. Infect. Immun 33:758-762; Riley, E.M., G. Anderson, L.N. Otoo, S. Jepsen, and B.M. Greenwood. 1988. Cellular immune responses to Plasmodium falciparum antigens in Gambian children during and after acute attack of falciparum malaria. Clin. Exp. Immunol 73:17-22; Ho, M., H.K. Webster, B. Green, S. Looareesuwan, S. Kongchareon, and N.J. White. 1988. Defective production of and response to IL-2 in acute human falciparum malaria. J. Immunol. 141:2755-2759.) that prevented the expansion of antigen-specific clones. The stability of BJ use observed between groups was consistent with this observation (data not shown).

Modification of the TCRB repertoire was evidenced only in the PBL of CM$^+$ mice in contrast with the usually well-accepted idea that PBL reflects spleen. This is also in contrast with spleen being necessary for the occurrence of CM (Curfs, J.H., T.P. Schetters, C.C. Hermsen, C.R. Jerusalem, A.A. van Zon, and W.M. Eling. 1989. Immunological aspects of cerebral lesions in murine malaria. Clin. Exp. Immunol 75:136-140., Mercado, T.I. 1973. Plasmodium berghei. Inhibition by splenectomy of a paralyzing syndrome in infected rats. Exp. Parasitol 34:142-144; Hermsen, C.C., E. Mommers, T. van de Wiel, R.W. Sauerwein, and W.M. Eling. 1998. Convulsions due to increased permeability of the blood-brain barrier in experimental cerebral malaria can be prevented by splenectomy or anti-T cell treatment. J. Infect. Dis 178:1225-1227.). Absence of specific alteration in the spleen of CM^+ mice could be explained by the dilution of stimulated cells in the bulk of T cells present in this organ and the fact that they leave to recirculate when they were activated (Mackay, C.R., and U.H. von Andrian. 2001. Memory T cells: local heroes in the struggle for immunity. Science 291:2323-2324.).

T cell clones associated with neuropathogenesis can be of different types. First, they can be private, specific to one individual, or recurrent, present in different individuals and sometimes designated as public clones. Secondly, their function might be pathogenic, protective or regulatory. Recurrent clones associated to neuropathogenesis were identified by assessing their presence in CM^+ mice and absence in CM^- mice. Five out of the ten most recurrent and differentially expressed peaks use the BV8.1 segment. Observations that supported their having a pathogenic role is that depletion of $BV8.1/2^+$ cells diminished the incidence of CM from 90% to 40% (Boubou, M.I., A. Collette, D. Voegtlé, D. Mazier, P.-A. Cazenave, and S. Pied. 1999. T cell response in malaria pathogenesis: selective increase in T cells carrying the TCR Vb8 during experimental cerebral malaria. Int. Immunol 11:1553-1562.). Others peaks including the ones using BV6, BV7 and BV9 segments have been identified in CM^+ mice (Table 2). This relates to the absence of CM in mice treated with a superantigen that deleted BV6, 7, 8.1 and 9 using T cells (Boubou, M.I., A. Collette, D. Voegtlé, D. Mazier, P.-A. Cazenave, and S. Pied. 1999. T cell response in malaria pathogenesis: selective increase in T cells carrying the TCR Vb8 during experimental cerebral malaria. Int. Immunol 11:1553-1562., Gorgette, O., A. Existe, M. Idrissa-Boubou, S. Bagot, J.-L. Guénet, D. Mazier, P.-A. Cazenave, and S. Pied. 2002. Deletion of T cells bearing Vb8.1 TCR following MTV-7 integration confers resistance to murine cerebral malaria. Infect. Immun In press.).

No CM^--specific recurrent peak (data not shown) were identified, which could be involved in protection, but a BV2-BJ1.3 peak was recurrently present both in CM^+ and CM^- mice as judged by the high score obtained for the two groups (Table 2 and Figure 5). This peak was by far the most expanded peak in CM^- mice (data not shown). Mechanisms contributing to neuropathogenesis could thus be the result of a regulatory pathway between BV2 and BV8.1-expanded clones leading to alteration of their cytokine profiles (de Kossodo, S., and G.E. Grau. 1993. Profiles of cytokine production in relation with susceptibility to cerebral malaria. J. Immunol. 151:4811-4820.).

Altogether, these results may suggest that few T cell clones were implicated in the development of CM. The immunological history for each individual shapes the emergent repertoire differentially between inbred individuals (Bousso, P., A. Casrouge, J.D. Altman, M. Haury, J. Kanellopoulos, J.P. Abastado, and P. Kourilsky. 1998. Individual variations in the murine T Cell response to a specific peptide reflect variability in naive repertoires. Immunity 9:169-178.). These variations could explain why, among a group of genetically identical mice infected with the same stabilate of parasites, only some developed CM.

In a previous study, it was observed by flow cytometry an increase of BV8.1/2+ T cells in the PBL of CM^+ mice while no expansion was seen in the spleen of CM^+ mice nor in the PBL and spleen of CM^- mice ((Boubou, M.I., A. Collette, D. Voegtlé, D. Mazier, P.-A. Cazenave, and S. Pied. 1999. T cell response in malaria pathogenesis: selective increase in T cells carrying the TCR Vb8 during experimental cerebral malaria. Int. Immunol 11:1553-1562.); data not shown). This increase, if caused by the expansion of few T cell clones, should distort the bell-shaped CDR3 length distribution of BV8.1/2-BC profiles. However, the PBL BV8.1/2-BC repertoire of CM$^+$ mice was not perturbed by comparison with CTR mice (ANOVA, p=0.71) or CM$^-$ mice (ANOVA, p=0.29). In addition, no modification of the BJ segment used was observed in the CM$^+$ mice. The increase of the representation of BV8.1/2 cells in the PBL of CM$^+$ mice thus cannot be attributed to a mono- or oligo-clonal increase and is therefore polyclonal. BV8.1/2 could be stimulated by a superantigen-like molecule, as observed in *P. yoelii* infection (Pied, S., D. Voegtlé, M. Marussig, L. Rénia, F. Miltgen, D. Mazier, and P.-A. Cazenave. 1997. Evidence for a superantigenic activity during murine malaria infection. Int. Immunol 9:17-25.). Implications of such molecules in pathogenesis have been reported for infections by *Toxoplasma gondii* (Subauste, C.S., F. Fuh, R.D. Malefyt, and J.S. Remington. 1998. ab T cell response to Toxoplasma gondii in previously unexposed Individuals. J. Immunol. 160:3403-3411) and *Leishmania infantum* (A. Sassi, A. Collette *et al*, unpublished results).

The original method presented in this report allowed the exhaustive analysis of immune repertoires. Applied to a mouse model of malaria, it demonstrated that the neuropathology induced by *P. berghei* ANKA was associated with a global perturbation of TCRB repertoires specifically found in PBL together with the recurrent expansion of few T cell clones.

This method can easily be transposed to human malaria since PBL are easily accessible to experiment. It is intriguing to know if the same association between PBL perturbation and neuropathology can be found in *P. falciparum* malaria. Furthermore, classification experiments allowed separation of the CM$^+$ and CM$^-$ mice and thus provide new tools for a better understanding of the immune response during malaria in humans. These hypotheses are being tested by studying cohorts of malaria patients. Finally, the original approach for deciphering lymphocyte repertoires can be transposed to various pathological conditions. For instance, this methodology is used in clinical follow-ups of patients after bone marrow transplantation or vaccination. The results and approach presented provide a promising basis for the bioinformatics revolution in the field of immunology.

Example 2-1: TCRB repertoire data

Parasite, mice and TCR BV-BC and BV-BJ CDR3 spectratype raw data, used in the present study, are described in Example 1.1 above. Briefly, eight-week old B10.D2 mice were infected by intraperitoneal injection of 10^6 red blood cells parasitized by PbA. Between day 7 to 10 after infection, PbA induces in some mice a cerebral syndrome, which is used as a model for cerebral malaria. Mice that do not show any cerebral signs die of severe anemia due to hyperparasitemia three weeks after infection. These mice, designated HP^+ in this report, were sacrificed between day 11 to 16 after infection when exhibiting a parasitemia above 20%, indicating that they had survived the critical cerebral malaria period. The PBL and spleen TCRB repertoires of these mice were analyzed. Percentages of use of each CDR3 peak inside its profile were computed and assembled into a peak database with ISEApeaks DataAnalyser module.

Example 2-2: Data smoothing algorithm

A series of data filters were implemented in the DataSmoother module of the ISEApeaks product to obtain a CDR3 spectratype, which is composed of several peaks. The i^{th} peak of the j^{th} profile of the k^{th} sample can be described by its area $A_{i,j,k}$ in arbitrary unit and its nucleotide length $L_{i,j,k}$. The peaks are ordered by increasing length. For each profile, λ_j will represent the theoretical PCR product length for a CDR3 of 10 amino acids.

The first filter removed background noise and peaks inferior to a defined cut-off according to user-defined parameters. In addition, peaks with identical length are summed.

The second filter corrected peaks for which $L_{i+1,j,k}$-$L_{i,j,k}$ = 1, designated as adjacent peaks, and, thus, do not respect the three nucleotide spacing expected for in-frame V(D)J rearrangements. Accordingly, each pair of adjacent peaks were tested to determine whether $L_{i,j,k}$ - λ_j or $L_{i+1,j,k}$ - λ_j are a multiple of 3. When this criteria was met, the corresponding peak is attributed to the biologically expected peak and the adjacent peak is summed. Otherwise, it is not possible to decide since the peaks are between two consecutive expected peaks. Hence, they are left unmodified for manually corrected by the DataFormatter module of the ISEApeaks product.

The final filter solved more subtle problems.

Following the peak processing by the aforementioned filters, some peaks remained (designated as ambiguous peaks), which can be attributed to the same theoretical length; however, these ambiguous peaks, in fact, corresponded to two distinct theoretical peaks. To estimate a possible shift which causes this problem in a profile (j, k), the Euclidean division of $L_{i,j,k}$ - λ_j by 3 was calculated with the remainder being assigned to an element of {-1, 0, 1}. If the mean of remainders was superior to 0.5 (respectively inferior to -0.5), all profile's lengths were shifted by -1 nucleotide length (respectively 1). Afterwards, the profile was analyzed again to warn the user of remaining ambiguous peaks.

Example 2-3: Methods for CDR3 spectratype analyses and statistics

Three different methods were implemented in the MacOS ISEApeaks product to analyze the peak database.

In a first approach, multivariate statistics were used to give a global description of the peak database using the percentage of use of each CDR3 length directly. Each repertoire was considered as a vector in n-dimensional space, where n is the number of variables that describe the repertoires. Missing values were replaced by the overall mean of the variable as recommended in Rencher (Methods of multivariate analysis. 1995, J. Wiley, New York). The resulting number of variables (230 peaks for BV-BC repertoires) was too high for theoretical constraints of Discriminant Analysis. The Principal Component Analysis (PCA) was first used to reduce the number of variables. PCA extracts new variables from the data set that retains the variability contained in the original data set. Linear Discriminant Analysis (DA) was then performed on the new data set to compare the different groups. DA computes discriminant functions that maximize inter-group variation and minimize intra-group variation. Significance of each discriminant function was tested using χ^2 approximation of Wilks' statistics Rencher (Methods of multivariate analysis. 1995, J. Wiley, New York).

The second method estimated the perturbation of a BV-BC repertoire by comparison to a control group (Gorochov et al. *Nat. Med* 1998, 4;215 and Han et al., *J. Immunol.* 1999, 163;301). µDBV-BC was defined as the mean of BV-BC perturbations, which corresponds to the overall sample perturbation. Similarly, DBV-BJ was defined as the BV-BJ perturbations. µµDBV-BJ was defined as the overall BV-BJ perturbation for all BV-BJ combinations analyzed. All BV-BC and BV-BJ perturbations ranged from 0 (identical profiles) to 100 (completely different profiles). Two-way Analysis of Variance (ANOVA) and t-test were used to analyze sample µDBV-BC and µµDBV-BJ, respectively. When ANOVA was significant, comparisons between two categories were performed with Fischer's Protected Least Significant Difference. To further assess similarity between samples, k-mean clustering with the Euclidian distance and hierarchical clustering with Ward minimum variance criteria were used. Significance of the clusters obtained from k-mean clustering were assessed by computing the probability of observing by chance the number of samples of a particular group within the cluster (Tavazoie et al. *Nat. Genet.* 1999, 22;281).

In the third method, recurrent expansions in a group of samples were scored by OligoScore computed by ISEApeaks (Collette and Six, *Bioinformatics* 2002, 18; 329). A score was computed for each peak in each group of samples. OligoScore of the control group was then subtracted to remove background noise.

PCA, DA and k-mean clustering were performed with SYSTAT 10 (SPSS). ANOVA were performed with StatView 5.0 (SAS Institute Inc.). Theoretical considerations on multivariate statistics used here can be found in Rencher (Methods of multivariate analysis. 1995, J. Wiley, New York). Statistics were considered significant when $p<0.01$.

Example 2-4: ISEApeaks modules for CDR3 spectratyping analysis

To enable the analysis of large amount of CDR3 spectratype data, ISEApeaks product was employed to extract, format, and gather data from the automated sequencers GeneScan, Immunoscope or Genotester into an Excel peak database as already described (Collette and Six, *Bioinformatics* 2002, 18; 329). Two modules were specifically designed for CDR3 spectratype data, addressing the problem of data smoothing and data analysis (Figure 6).

Differences in lane migration and peak finding algorithm of GeneScan, Immunoscope or Genotester automated sequencers results in divergence between what one observes on a gel and the figures given by these softwares (Figure 7A and 2B). This problem is of paramount importance since it is not possible to build the peak database when two experimental peaks are found where only one is theoretically expected. An automatic analysis thus requires that this problem be first solved. CDR3 peaks should be separated by three nucleotides as expected for in-frame rearrangements found in peripheral lymphocytes. On this basis, filters were designed and implemented in the DataSmoothing module to smooth peak data (Figure 7C).

The resolution of peak data problems were assessed with 9 mouse BV-BC repertoires. Raw data presented 141 problems of which one third were solved by the first filter consisting in background noise removal and addition of peaks with identical length. 82 % of the 45 adjacent peaks and 60 % of the 50 ambiguous peaks were resolved with the last two filters, thereby considerably reducing the amount of time necessary for manual data correction. Altogether, 80% of peak problems were solved. The remaining inconsistencies, highlighted in ISEApeaks DataFormatter sheets, needed to be corrected by the user by comparing two CDR3 profiles of the same BV-BC or BV-BJ combination for two different samples.

The second module, DataAnalyser, module was then used to build and analyze the CDR3 peak database. ISEApeaks enabled the assessment of repertoire perturbations (Gorochov et al. *Nat. Med* 1998, 4;215 and Dechanet et al. *J. Clin. Invest.* 1999, 103, 1437), oligoclonal expansions (Cochet et al., *Eur. J. Immunol.* 1992, 22; 2639) or recurrent clones defined as rearrangements that were used by different individuals in response to a particular infection or immunization (Collette and Six, *Bioinformatics* 2002, 18; 329). The average CDR3 length of each profile, the average repertoire for each group and the peak number per profile were also be computed. In addition, a summary table representing the different samples in which diversity of each BV-BC or BV-BJ profiles was plotted, which was color-coded. Altogether, DataAnalyser module enabled a global and objective analysis of immune repertoire data.

In a third module, tools were devised to facilitate the use of the Immunoscope products. For example, a first tool created Immunoscope 3.1α analysis macros using DataParameter files. A second tool tackled the problem of data representation. Typically, experimenters need to gather CDR3 spectratypes of different samples, which resulted in files containing CDR3 profiles being scattered in different folders with abstruse names. ISEApeaks solves this problem by assigning a file, which describes each load of a run. DataExtractor then uses this parameter file to find, copy and rename all needed files before assembling them in a template provided in the product of up to 16 12 profiles.

Example 2-5: Analysis of BV-BC repertoire modifications during mouse malaria

The TCRB repertoire in a group of PbA-infected mice, developing severe anemia due to hyperparasitemia (HP), was analyzed to study repertoire modifications induced by PbA infection by comparing HP mice to control (CTR) non-infected mice. Multivariate statistics were required to apprehend the complex changes induced by this parasite.

BV-BC repertoire was analyzed both in the PBL and spleen compartments, representing 32 samples. ISEApeaks was used to retrieve raw data from the 672 CDR3 profiles. After smoothing, the peak database was formed. In a first approach, the differences between the four groups of samples (CTR PBL, CTR spleen, HP PBL and HP spleen) was assessed by Discriminant Analysis (DA) after reduction of the number of CDR3 peak variables by Principal Components Analysis (PCA). Results of DA performed on this new data set, which retained 99% of the original data information, are displayed in Figure 8. Only the first two discriminant functions are statistically significant indicating that only three groups could be separated: HP PBL, HP spleen and CTR samples. Thus, infection by PbA globally induced alteration of the BV-BC repertoires.

To get insights into the nature of this alteration, indexes of perturbation (DBV-BC) between samples and the CTR spleen group were computed. DBV-BC data was visualized in an objectively color-coded array (Figure 9). Globally, perturbation in HP samples seemed greater. To test this observation, sample perturbations (μDBV-BC), the mean of the DBV-BC, were compared by Analysis of Variance (ANOVA) using the organ (PBL or spleen) and clinical status (CTR or HP) factors (Figure 10A). Only the clinical status factor was significant ($p<0.0001$), confirming that the infection induces an alteration of the TCRB repertoire without distinction of blood versus spleen compartments.

Typically, both DA on peak data and ANOVA on perturbation data are used to assess the difference between known groups of samples. However, k-mean clustering, another multivariate statistical method, can be used to cluster related samples without *a priori* knowledge of group composition. This method was applied on BV-BC repertoire data (Figure 10B). Cluster analysis allowed separation of all CTR samples and infected samples but one ($p=2.8 \times 10^{-8}$). Thus, three independent statistical methods showed that infection by PbA induced an alteration of the BV-BC repertoire both in the blood and the spleen compartments.

Example 2-6: Analysis of blood BV-BJ repertoire modifications during mouse malaria

The precision of the repertoire description by BV-BC CDR3 spectratypes was assessed by studying BV-BJ repertoires. The BV-BJ spectratypes of 9 (out 21) BV genes, representing more than two third of the TCRB repertoire, were studied in the PBL compartment. Likewise, the repertoire alteration of each sample as compared to the CTR PBL reference group was studied. ISEApeaks computed the perturbation for each BV and BJ segments (DBV-BJ). The average of DBV-BJ among BV and BJ segments (μμDBV-BJ) quantitatively measured the alteration for each sample. PBL BV-BJ repertoires of HP mice (mean=24.5) were also more altered than in CTR mice (mean=11.9; t-test, df=9, p=0.0001).

To assess whether this alteration of BV-BJ repertoires could also enable the clustering of samples without *a priori* knowledge of group composition, hierarchical and k-mean clustering were used. Both methods achieved a complete separation of groups (p=0.002), confirming a strong dissimilarity between these samples (Figure 11).

Example 2-7: Identification of recurrent clones during malaria

Another interesting feature of CDR3 repertoire analyzed was the finding of recurrent CDR3 expansions of similar clones responding to challenge. The PbA parasite contained a very large number of antigens and induced a profound alteration of the TCRB repertoire. OligoScore, which objectively scores each of the 1056 peaks for their recurrence (Collette and Six, *Bioinformatics* 2002, 18; 329) was employed. No peak was found recurrently expanded in the BV-BC repertoires. Analysis of BV-BJ data enabled the finding of recurrent peaks, particularly of BV2-BJ1.3 and BV2-BJ1.1 peak with a CDR3 length of nine amino-acids (Table 4).

Table 4: Identification of BV-BJ recurrent peaks during experimental malaria.

Rank	Description	CDR3	OligoScores		Number[2]		□HP/CTR
		(aa)[1]	HP	CTR	HP	CTR	
1	BV2-BJ1.3	9	6.12	0.25	4	5	5.87
2	BV9-BJ1.5	10	3.46	0.01	1	1	3.45
3	BV2-BJ1.1	9	1.37	0.25	4	6	1.12
4	BV8.2-BJ2.2	10	1.24	0.19	5	6	1.05
5	BV7-BJ1.2	7	1.17	0.13	4	6	1.04
6	BV6-BJ1.3	10	1.09	0.19	3	6	0.90
7	BV7-BJ1.6	9	1.58	0.68	2	5	0.90
8	BV4-BJ1.3	10	1.48	0.59	4	6	0.89
9	BV6-BJ2.1	9	0.89	0.07	5	6	0.82
10	BV2-BJ1.3	10	1.19	0.37	4	5	0.82

The 1056 peaks of BV-BJ repertoires were scored for recurrent expansion in the CTR and HP group by OligoScore (Collette and Six, *Bioinformatics* 2002, 18; 329). Peaks were then sorted decreasingly according to the difference (□HT/CTR) between the score of the HP group and the one of the CTR group. All OligoScores can be obtained on the supplemental data web page.

[1] CDR3 lengths are indicated in amino-acids.

[2] Number of samples for which the given profile data was analyzable is indicated. A total of 4 CTR and 6 HP mice were studied.

Figure 12 shows the corresponding BV2-BJ1.3 and BV2-BJ1.1 CDR3 profiles. BV9-BJ1.5 cannot be considered recurrently expanded since sufficient signal was only obtained for one sample in each group. Direct sequencing confirmed that these expansions represented public clones since very similar CDR3 sequences were found (Table 5).

Table 5: CDR3 direct sequencing of identified PBL expansions in infected mice.

TCRBJ	CDR3	Mice	<- CDR3[1] Sequences ->				
	(aa)		TCRBV2 ->			<-	TCRBJ
1.3	9	#1	CTC	S	XTG	SGNTL	YFG
		#2	CTC	S	SAK	SGNTL	YFG
		#3	CTC	S	SGT	SGNTL	YFG
1.1	9	#1	CTC	S	GTGA	NTEV	FFG
		#2	CTC	S	GTGA	NTEV	FFG
		#3	CTC	S	XTGA	NTEV	FFG
		#4	CTC	S	GTGA	NTEV	FFG

BV2-BJ PCR products were directly sequenced using both BV- and BJ-specific primers.

[1] X stands for a position that could not be determined due to ambiguous reading of the nucleotide sequence. Following (Kabat et al, *Sequences of proteins of immunological interest* 1991, Bethesda, MD) the CDR3 region was taken as encompassing amino-acids 95 to 106.

Example 3.1 ISEApeaks package 2.0.1 and its interface with GeneScan and Immunoscope to analyze immune repertoires-Overview

The ISEApeaks 2.0.1α product can be used with all versions of the immunoscope products (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.). However, using the latest one, Immunoscope 3.1α, will allow the use of all ISEApeaks utilities devised for this product (see below). Hence, fluorescent signal from any Sequencing Automate of PE Biosystems (370A, 373, 310 or 377) can be analysed after Immunoscope or GeneScan 2.2.1 analysis. Similar results can be achieved by using other automated sequencers with the appropriate software, such as MEGABACE sequencer AND GENOTESTER^{TM} software. The Immunoscope technique (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323.) with or without utilizing Excel, Immunoscope and GeneScan softwares can be used with the ISEApeaks 2.0.1α product.

Preceding the use of the ISEApeaks product, the different steps of the analysis are summarized in Figure 13. Samples are amplified using V to C or V to J primer combinations and PCR products are labelled by run-off amplifications with fluorescent primers. Labelled PCR products are loaded on an automatic sequencer to be separated according to their nucleotidic length. PE Biosystems product, Sequencing Analysis, is used to generate raw data files, one per well. Immunoscope and GeneScan use these files to analyse the fluorescent signal: they both generate files containing nucleotidic sizes and areas of identified peaks. Through Immunoscope macros, it is possible to automatically analyse the whole gel, generating as many PICT files as the number of well and the number of loads per well. These PICT files contain the profile of the analysed fluorescent signal as well as the length and area of the peaks. They can be assembled to form the image of a whole repertoire using template files. A "profile" will designate either the analysed fluorescent signal stored in a PICT file or the description of the CDR3 peaks by their nucleotidic size and peak area.

The ISEApeaks product has been developed to solve the following example of problems that may arise (Collette, A. et al. 2002. High throughput screening and analysis for CDR3 spectratypes during malaria with the ISEApeaks software package. manuscript in preparation; Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.). First, no tool is available in Immunoscope or GeneScan to retrieve peak data. Secondly, there is a discrepancy between the number of peaks as seen by eye and the number of peaks given in the Immunoscope PICT or GeneScan files. Thirdly, there is no product to address the problem of data analysis while the amount of data generated is increasing with the use of high throughput sequencers. ISEApeaks gives answers to these problems. Figure 13 shows the overall architecture of the ISEApeaks product. Data generated by Immunoscope or GeneScan can be extracted using DataExtractor. DataSmoother is suited for immune repertoire application. DataExtractor and DataSmoother are parameterised with files created with DataParameter. Data obtained can then be imported into Excel files using DataFormatter Excel templates and the ISEApeaks Excel Add-in. After using DataAnalyser and the add-in, the data peaks of a whole experiment with different organs, individuals and groups, can be gathered. Several tests are implemented to perform repertoire perturbation analyses.

ISEApeaks package helps the experimenter to extract and gather peaks in an Excel database. This database is used to analyse immune repertoires. Futher, specific modules have been implemented (i.e. DataSmoother and the analysis part of DataAnalyser). DataParameter enables the user to parameterize the extraction and smoothing of data using DataExtractor and DataSmoother.

DataFormatter displays the data in Excel. DataAnalyser performs both the gathering of peaks scattered in different DataFormatter files and the analysis of this peak database for immune repertoires. Useful utilities have been incorporated in the package such as automatic construction of Immunoscope macros or assembling of Immunoscope PICT files in a unique sheet. ISEApeaks softwares are displayed in underlined red.

DataFormatter is used to display the data in Excel. DataAnalyser performs both the gathering of peaks scattered in different DataFormatter files and the analysis of this peak database for immune repertoires. Useful utilities have been incorporated in the package such as automatic construction of Immunoscope macros or assembling of Immunoscope PICT files in a unique sheet. ISEApeaks softwares are displayed in underlined red.

ISEApeaks Excel add-in displays a menu bar from which one can launch macros after selecting the proper file type one is using. Clicking the right icon will launch the ISEApeaks Excel main menu, or use the shortcut (alt+option+m) (Figure 14) where one can access Preferences menu (Figure 15), create new files, get preferences or information about the Excel active file or reset the ISEApeaks menu bar.

ISEApeaks Excel add-in preferences are displayed in Figure 15. When screen updating is off, some long ISEApeaks macros will not update the screen so that computations are more rapid. Excel status bar (bottom of the screen) will indicate to user the status of the execution. kMaxPeakNb is the maximum number of peaks per profile that will be displayed (see DataFormatter section). kGelWellNb is the number of well of the sequencing gel one uses. For a 36-well automatic sequencer this preference value should be set to 36. kProfileNbPerRep is the number of profiles in a repertoire. This value is used to separate data in a gel. For instance, a common human Vβ-Cβ repertoire analyses 24 different combinations; hence this preference value should be set to 24; 36-well gel data will be displayed in 3 sheets, one per repertoire (double loading is assumed). Note that kGelWellNb and kProfileNbPerRep parameters can be modified from the ISEApeaks menu bar (see Figure 16) using the last button command: the first figure is the kGelWellNb and the second kProfileNbPerRep. kMaxColNb and kMaxLinNb is used by ISEApeaks in the assembling of Immunoscope profiles (see below) to determine how many files are to be assembled.

The ISEApeaks preferences menu enables the user to modify preference values used by the ISEApeaks XL add-in.

DataExtractor and DataSmoother are standard Macintosh applications. A menu indicates the different functionalities available. Dialog boxes prompt the user to chose parameter files or folders where data to be extracted are stored.

Different files are used by the ISEApeaks package. To help the user distinguish between files, they have been assigned different file types which are described below.

'APPL' files are ISEApeaks application file.

'CGEL' file are parameter files used by DataExtractor and DataSmoother to extract and smooth Immunoscope or GeneScan data of a whole gel. A CGEL file is also needed to create an Immunoscope 3.1α macro using the Utilities menu of DataExtractor. CGEL files are created with DataParameter.

'CPIC' file (CPICTPLACES files) are parameter files to use with DataExtractor to prepare the assembling of Immunoscope PICT files using Immunoscope. When having loaded different gels to analyse different samples one might want to gather a particular profile for all the analysed samples. To assist one with this tedious work, DataExtractor will automatically copy and rename the files corresponding to this particular profile as indicated by a CPICTPLACES parameter file. CPICTPLACES files are created with DataParameter. After DataExtractor has copied the desired PICT files, one can use the Immunoscope software to assemble the profile saved in the PICT files in the template provided by the ISEApeaks package. This template contains an array of empty profiles of 16 lines by 12 columns.

'DATA' file are output files of DataExtractor and DataSmoother containing peak data. These files will be used by DataFormatter to import the peaks in Excel.

'COUP' file are output log files of DataExtractor and DataSmoother that can be saved.

'TEXT' file are usual TEXT files that contain the Immunoscope 3.1α macros created with DataExtractor.

'PREF' files are ISEApeaks preferences files.

Example 3.2: DataParameter

a) CGEL parameter files

DataParameter will help one to create the CGEL parameter file needed for DataExtractor and DataSmoother. In this file, one will give information concerning the nature of what one wants to extract and analyse and how one wants to get it done (Figure 17). In this example, the selected Excel worksheet shows a part of CGEL parameter file data.

Open a DataParameter file using Excel® 98. Figure 17 shows an example of a DataParameter file. For each well of the sequencing gel, one can specify at least one of the following different parameters.

The 'mIsConsidered' parameter: setting the parameter to 0 means one does not want to analyse this well, for instance because no sample has been loaded in the well. Setting it to 1 indicates that this well will be extracted and smoothed.

The 'mDescription' parameter: in this box, one can put a string of characters that depicts the nature of the sample one loaded in this well. This string must not contain the ';' character.

The 'mLength' parameter: put in this box the size of one's expected fragment. For instance in a repertoire analysis, put the value expected for a PCR fragment with a 10 aminoacids CDR3 .

The 'mNewOrder' parameter: this parameter sets the new order of appearance of the different wells. This function enables automatic sorting of the different wells .This new order will be used for extraction and in the 3e) paragraph.

Double loading is a widespread way to extend the loading capacity of a sequencing gel. That's why double loading is possible in DataParameter. If one doesn't double load, just set to 0 the mIsConsidered parameter of all the wells of the second load. The other parameters in the left corner of the worksheet are the general parameters for the analysis of the run. They are as follows.

The 'mTypicalPictFileName' parameter is the name of one of the data files. Note that the 'mTypicalPictFileName' parameter must contain no other figures than the two figures automatically attributed by the Immunoscope software or created by ISEApeaks from GeneScan data. This restriction makes impossible the use of name such as 'Z 24/2/98 24.1: use instead 'Z 24.1'. This string must not contain the ';' character.

The 'mGelWellNb' parameter: the well number of the sequencing gel one loaded. This value can be for instance 36, 48 or 96.

The 'mBackgroundNoise' value is the threshold under which signal is considered to be background and will be ignored.

The 'mCutoff' value is used by DataSmoother: it is the percentage of the maximal area under which one wants to ignore the signal.

The ‘mMethodFlag’ value will be discussed in the DataSmoother paragaraph. Set this value to 2 to use a correction based on the value of mTheoricLength. Set this value to 3 to use a correction based on the length of the peak of maximal area of the profile.

In the ISEApeaks menu bar, one can access the different macros of DataParameter. The 'ImportCGelFile' macro can be used to import a parameter file in the Excel file where parameters can be easily modified. A dialog box will prompt the user to choose a file. The 'CreateCGelFile' macro can be used to create a parameter file. Note that if one is doing the same experiments (loading each time one's gels in the same manner), one will need only once to create a CGEL parameter file. The 'CGELPageSetUp' macro will create an empty sheet containing a page set up that can be used to enter new values for a CGEL file. To find about the use of the CGEL parameter files, please see the DataExtractor or DataSmoother sections.

Finally, executing the ImportCGelFile or CreateCGelFile macro will update automatically the Excel preferences with the values used in the file imported or created. When using the CreateCGelFile macro, it might happen that Excel seems to be frozen. To solve this problem, see the Pitfalls section below.

b) CPICTPLACES parameter files

The ISEApeaks package provides a useful utility to solve the tedious problem of Immunoscope PICT files assembling. Let's take an example: say one has analysed the Vβ-Cβ repertoire (24 profiles per sample) of 6 mice and the Vβ8.1-Jβ repertoires (12 profiles per sample) of 6 mice. One surely wants to compare the Vβ8.1-Cβ and the 12 Vβ8.1-Jβ profiles of these 6 mice, but all these profiles are stored in 13*6 = 78 different files scattered in different folders. To put these 78 profiles on a unique image, one will have to open each of these files and copy their profile to a unique document. The ISEApeaks package provides an easy solution. With CPICTPLACES parameter files, one can specify the PICT files one wants to assemble. These files are copied by DataExtractor utility. Finally, the copied files are assembled by Immunoscope software using a template provided in the ISEApeaks package or one that one customised.

DataParameter helps one to create the CPICTPLACES parameter files needed for preparing the assembling of PICT files scattered in different folders. In this file, one will provide information concerning the PICT files one wants to gather on a same paper sheet to compare them. The DataExtractor section will describe the use of CPICTPLACES parameter files to prepare the assembling of PICT files. Figure 18 shows an example of a DataParameter file. On the top left part of the sheet, the general parameters needed for all of the assembling procedure are displayed. There parameters are explained below.

The 'mDestFolderName' parameter specifies the name of the folder where duplicated PICT files will be stored.

The 'mMaxColNb' parameter specifies the number of rows used in one's repertoire template file.

The 'mMaxLinNb' parameter specifies the number of lines used in one's repertoire template file.

In this example, one can see a portion of a CPICTPLACES worksheet. For each file, different parameters are listed that will enable DataExtractor utility to process through the different gel folders and select the PICT files indicated in the CPICTPLACES parameter file. This example corresponds to the case envisaged in the text (assembly of Vβ8.1-Cβ and Vβ8.1-Jβ for 6 mice, see also the example folder).

For each file, one can specify different parameters. These parameters are described below.

The 'mIsConsidered' parameter: setting the parameter to 0 means one does not want to use this profile position. Setting it to 1 implies that this profile position will be considered.

The 'mFolderName' parameter is the name of the folder where the PICT file one wants to put in the corresponding profile is located. This string must not contain the ';' character.

The 'mCGELFileName' parameter is the name of the CGEL parameter file of the corresponding mFolderName. This CGEL file describes how the gel whose data are stored in this folder was loaded and must be in the same folder as the PICT files. Note that this requires that the CGEL parameter is systematically left in the folder. This string must not contain the ';' character.

The 'mSet' parameter is the repertoire set number (each set has as defined mWellsNbPerSet wells). For instance, three mice Vβ-Cβ repertoires can be loaded on a same gel, while one can load the Vβ-Jβ repertoires of 6 different Vβ on a gel. In the first example, mSet should be set to 3 if one wants to take profiles of the third repertoire while in the second one mSet should be set to 1 as there is only one occurrence of a particular Vβ-Jβ combination since the 6 Vβ are different. This parameter should never be set to 0.

The 'mWellsNbPerSet' parameter is the number of profiles per repertoire. For a double-loaded 36-well gel used to analyse 3 Vβ-Cβ repertoires, this parameter should be set to 24. For a double-loaded 36-well gel used to analyse the Vβ-Jβ repertoires of 6 different Vβ, this parameter should be set to 72 as all profile description will be different. This parameter should never be set to 0.

The 'mDescription' parameter is the description of the nature of the analysed profile one wants DataExtractor to pick among all the Pict files of this gel. This string must not contain the ';' character.

Note that these parameters should be set in accordance with the parameters given in the corresponding CGEL parameter files, especially for the mDescription and that each set should contain only one occurrence of each mDescription! Using the ISEApeaks menu bar one can access to the 2 macros proposed to deal with CPICTPLACES parameter files. The 'ImportCPictPlacesFile' macro allows one to import a CPICTPLACES parameter file. A dialog box will prompt the user to choose the file. The 'CreateCPictPlacesFile' macro will use the data of the active Excel sheet to create a new CPICTPLACES parameter file. The CPICTPLACESPageSetUp macro enables the user to build an empty formatted sheet to be filled with new parameters. Finally, importation or creation of CPICPLACES files will update the Excel Preferences of the add-in with the values used in the file.

The use of CPICTPLACES parameter files is described in detail in the DataExtractor section. The provided template allows the assembly of up to 192 profiles organised in 16 lines by 12 rows.

Example 3.3: DataExtractor

DataExtractor extracts the data from Immunoscope, GeneScan, and Genotester files using the information provided by a CGEL parameter file (Figure 13). GeneScan data must first be exported by the use of the GSExportscript of the ISEApeaks package. After detailing the extraction process for both Immunoscope and GeneScan data, 3 Immunoscope-related utilities are described: renaming of the PICT files, creation of Immunoscope 3.1α macros using CGEL parameter file and assembling of PICT files in a repertoire template to gather profiles analysed on different gels.

a) GeneScan export script

Gel data are stored in GeneScan files, one per well. The first task to perform is to export data peaks to text files readable by ISEApeaks. To assist one, ISEApeaks provides an applescript to automate the exportation process. Use DataParameter to generate the list of GS Filename: names should be entered just below the mNewOrder parameter (see Figure 18). Launch the GSCreateFileList macro using the ISEApeaks Excel menu bar. The provided apple script will use this file.

Then open GeneScan to set the current folder of GeneScan to the folder where one's GeneScan files are stored: a common way to do this is to open one of the sample files and cancel the action when in the proper folder. Launch the apple script GSExportscript, choose the data-containing folder, select the file names list and, finally, indicate the first number to use in exported files. The script will open each GeneScan files and export the data. The subsequent step of data extraction is described in the next section.

b) Data extraction

DataExtractor reads the data files and extracts the length and area of the peaks either of Immunoscope PICT files or exported GeneScan data files or Genotester exported data (see previous section). Launch the DataExtractor program by double clicking on the file icon (for Immunoscope data only). Choose the appropriate function in the menu. DataExtractor can extract data for all well formats and ensures the reordering of the samples (indicated in the CGEL parameter file).

To Launch the extraction of Immunoscope data, select the 'For Immunoscope Data' item in the 'Extraction' menu (shortcut -I). To Launch the extraction of GeneScan data files exported with ISEApeaks (see GeneScan export script section), select the 'For GeneScan Data' item in the 'Extraction' menu (shortcut -G). To Launch the extraction of MegaBACE data file , i.e. Genotester datafile, exported in tab-delimited TEXT with Excel, select the 'For MegaBACE Data' item in the 'Extraction' menu (shortcut -M). The user is then asked by a dialog box to select a 'CGEL' parameter file. This file must be in the same folder as the data files: it is very convenient to leave the parameter file used to extract the data in the gel folder, in case one wants to look back to what has been done or use PICT file assembling utility. During the execution, the status of the program is shown in a window. This output file can be saved as a text file and lists extracted files. Extracted files are listed as well as adjacent and ambiguous peaks, which could not be resolved by smoothing filters. The program generates a 'data.0' file. This file contains all the data stored in data files, without any computing which can also be imported in DataFormatter (see below page 75). At the moment, DataExtractor only keeps peaks whose lengths are +/-20 nt of the mTheoriclength value.

c) Immunoscope utilities: renaming ".Pict" files.

Some useful utilities for Immunoscope© have been incorporated in the DataExtractor program: one can access them through the 'Utilities' menu. Use the 'rename .Pict files' utility to modify the names of one's Immunoscope PICT of exported GeneScan files: one will be asked to select the old first well file, to indicate the desired new name and the gel well number. This can be useful when the files one has generated are named 'Z XX.Pict.3' by Immunoscope: one surely wants them to be renamed 'Z XX.Pict.1' to be able to assemble them in one's own template repertoire!

d) Immunoscope utilities: creating Immunoscope 3.1α macros.

The second utility enables one to create Immunoscope© 3.1α macro using the parameter file created with DataParameter. This utility is not compatible with older versions of Immunoscope. One will be asked to select a CGEL parameter file, to indicate the range, a template Immunoscope macro that will be used to create the new macro and, finally, the location and name of the new macro to create. The Immunoscope template macro can be the one provided with the package (named 'Macrotemplate' for 36-well gels) or any Immunoscope 3.1α macro.

This useful utility will spare one the tedious work of modifying the macros by selecting each of the 72 (for a double-loaded 36-well gel) macro instructions to change the length between which Immunoscope should analyse the fluorescent signal. These operations are often a source of errors.

e) Immunoscope utilities: preparing the assembling of .Pict file profiles

The DataParameter section has dealt with the creation and importation of CPICTPLACES parameter files as well as its usefulness for Immunoscope data. Choose the

'Preparing .Pict file assembling' in the 'Utilities' menu. A dialog box will first ask for a CPICTPLACES parameter file, and then to select a folder containing all the data, that is the common root folder for all the data-containing folders. Finally, indicate a default PICT file to use. This default file will be used when mIsConsidered is set to 0 (see page 69): Immunoscope needs to find a PICT file for each profile of the template repertoire. The process status of the operation is shown: DataExtractor reads the parameter file, locates the proper files and renames them to the output folder. The output folder is created in the selected folder. Note that, when renaming, existing files are overwritten.

Now, form the image with created file set using Immunoscope and the '16x12 RepIS template' template file provided with the ISEApeaks package. Open Immunoscope, open the template file using the 'Open repertoire' option and choose the 'Construct Rep.' option. Select one of the files created by DataExtractor in the output folder. Immunoscope then puts each profile in the proper location.

Example 3.4: DataSmoother

DataSmoother does the smoothing of the peaks: suppression of background noise and of double peaks and other problem solving.

Open DataSmoother. Choose the 'To import and analyse the data' option, select the CGEL parameter file and a 'data.0' file created by DataExtractor. A window will describe the operations performed. Let's describe what one sees. DataSmoother first imports the data: for each profile 'mDescription' and 'mTheoricLength' parameters as well as the number of peaks imported are listed. Then the analysis starts: background is removed using the value one specified in the CGEL parameter file. Peaks which areas are less that mCutoff percent of the maximal area of the profile are deleted. After this step, the file 'data.1' is saved containing all gel peaks.

The two other filters have been specifically designed for immune repertoire data whose peaks should be spaced by 3 nt at the periphery, as required for in-frame rearrangements. Two different algorithms are possible for "adjacent" peaks correction. One of the adjacent peaks is summed to the other one according to the method chosen with the mMethodFlag. If the parameter mMethodFlag is set to 2, the reference will be the theoretical length calculated using mTheoricLength. When set to 3, the reference will be the length of the maximal area peak. After this step, the file 'data.2' is generated.

Then, DataSmoother tries to solve another problem: two consecutive peaks can sometimes be attributed to the same theoretical length. Let's call them "ambiguous" peaks. This problem arises because of a too imprecise length analysis by GeneScan or Immunoscope. This problem is crucial as one of the two peaks will be erased when trying to gather all the peaks of different samples using DataAnalyser. For each profile, different parameters are listed. Just track the profile for which one of this problem is found, one can see that DataSmoother tries to resolve it. After this try, DataSmoother searches again for ambiguous peaks: if some are still detected, the user will be prompted to solve the problem in DataFormatter. Finally, the file 'data.3' is generated. Don't forget to save the log of the DataSmoother program.

Example 3.5: ISEApeaks DES

ISEApeaks DES gathers DataExtractor and DataSmoother functions: one just runs one application instead of two and the 'data.0' file is automatically used to do the smoothing.

Example 3.6: DataFormatter

DataFormatter is used to display in Excel the data extracted and smoothed by DataExtractor and DataSmoother. In Excel, corrections can be easily done if needed. It is also a good start to perform custom analysis. Data of a sequencing gel is stored in a unique DataFormatter file. Data will be displayed according to the preferences values: data will be separated in kGelWellNb/kProfileNbPerRep different sheets (See CGEL parameter files section). If the DataFormatter file used does not contain enough data sheets (Figure 19), ISEApeaks will create missing sheets. For instance, human and mice Vβ-Cβ repertoires are usually composed of 24 profiles (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323., Even, J., A. Lim, I. Puisieux, L. Ferradini, P.Y. Dietrich, A. Toubert, T. Hercend, F. Triebel, C. Pannetier, and P. Kourilsky. 1995. T-cell repertoires in healthy and diseased human tissues analysed by T-cell receptor beta-chain CDR3 size determination: evidence for oligoclonal expansions in tumours and inflammatory diseases. Res. Immunol 146:65-80.). So one may have up to 3 different sets of 24 profiles per gel. The way these 3 sets will be filled depends on the CGEL parameter file one gave to extract the gel data

To import the data, use the 'ImportData' macro of the ISEApeaks menu bar. Select the 'data' file one wants to import. The 'Calculation' (alt+ option +c) macro is automatically launched: it calculates different percentages based on the imported data. Two graphical displays represent the obtained percentages. The 'LengthCheck' (alt+ option +l) macro is also launched automatically: consecutive peaks are highlighted in yellow, "ambiguous" peaks are highlighted in red, maximum peak in green and the first peak that can be attributed to the mTheoricLength in bold. Importation is dependent of the way mNewOrder orders were set in the CGEL parameter and also of the Preferences values. In particular, the number of different data sheets will be determined by the values of kGelWellNb and kProfileNbPerRep.

Maximal peaks are highlighted in green. On this examples, ISEApeaks preferences have been set to a 36-well gel and 12 profiles per repertoire (6 data sheets for 6 mouse V□ complete Vβ-Jβ repertoires).

One can use the 'mIsConsidered' parameter to sum a given peak to the previous one (mIsConsidered=2), or to the next one (mIsConsidered=3), or to exclude it (mIsConsidered=0). Cell background can be set to blue to keep track of modifications of original data (alt+option+b) or to blank (alt+option+e). Use these features to correct the adjacent or ambiguous peaks reported in yellow or in red. After finishing these modifications, run again the 'Calculation' macro to update calculations. Rerun the 'LengthCheck' macro which will warn the user if some lengths have not been corrected. Note that Calculation/LengthCheck applies only to the active worksheet. The 'PeakPageSetUp' and 'PercentPageSetUp' macros allow one to draw the display using preferences values of the peaks array and percentage, respectively.

Calculations are the percentage of use of each fragment for each combination (for instance, V to C or V to J) in the whole sample, percentage of use of each profile (mDescription) and percentage of use of each fragment length in each profile. These percentages are displayed in histograms ("%Profile" and "%CDR3" sheets).

Example 3.7: DataAnalyser

DataAnalyser has two functions: gathering the data stored in the different DataFormatter files and analysis of this large amount of data for immune repertoires. DataAnalyser macros can be accessed using the ISEApeaks menu bar. First will be described the parameterisation of DataAnalyser and the retrieval of data scattered in different DataFormatter files to form the peak database. Finally, all implemented perturbation and oligoclonality tests as well as some useful utilities will be detailed. Of course, DataAnalyser will use the preferences values of the ISEApeaks Excel Add-in.

a) Parameterisation

Open a DataAnalyser file. First one will need to specify the different files one wants to import. Choose the 'para' worksheet. Specify the name of the DataFormatter files where each set of data is to be found, the name of the worksheet, the group it belongs to and finally characters to depict what this sample is (Figure 20). Note that all the DataFormatter files must be in the same folder.

Here is described an experiment where 9 mice were analysed in a Vβ-Cβ repertoire. Two groups have been created to compare the repertoire in the two analysed mice groups (cf. the 'Group' column). Note that the group numbers must be consecutive integer values starting with 1.

b) The 'TakePeaks' macro

After parameterisation has been done, begin the retrieving and alignment of peaks by selecting the 'TakePeaks' item of the ISEApeaks menu bar. First choose if one wants to retrieve peaks' areas or percentage of use of peaks. Each DataFormatter file is opened and DataAnalyser gathers the data of same nature and length in the 'Peaks' sheet. For instance, in a repertoire analysis, all the percentage of use concerning the peak using TCRBV8.1 with a certain CDR3 length will be gathered (Figure 21). Peaks are sorted in order of appearance. ‘length failed’ means this particular length is not present in the sample. 'excluded' means that this profile is not considered for this particular repertoire. The 'LengthCheck' procedure (see below) is automatically launched to check that the lengths are correct. Status of the execution is indicated in the Excel status bar (bottom of the screen).

The parameters shown in Figure 20 have been used to import the corresponding data. For each peak, the mDescription, the length in amino-acids of the CDR3 region, the length of the PCR product in nucleotides and the length of the 10 amino-acids CDR3 PCR product in nucleotides are indicated. Followed for each sample by the length of the peak and the percentage of use of this peak among this profile.

c) The 'LengthCheck' macro

The procedure 'LengthCheck' is automatically launched after 'TakePeaks' macro to check that the lengths are correct. Especially, if all the ambiguous peaks reported by DataSmoother were not corrected in DataFormatter, some peaks will be missing and the sum of the areas of the different peaks for the profile will not be 100%! Consecutive peak lengths are highlighted in yellow. These problems need to be solved before going further, usually it implies to go back to the corresponding DataFormatter file. Status of the execution is indicated in the Excel status bar (bottom of the screen).

d) The 'ImportPercent' macro

The procedure 'ImportPercent' can be used to retrieve the percentage of use and the fluorescent signal of the profile. Figure 22 shows an example of the gathering of the percentage of use and signal, using 'ImportPercent'. The 'ImportPercent' procedure uses the parameters entered in the 'para' sheet, as for the 'TakePeaks' macro. Status of the execution is indicated in the Excel status bar (bottom of the screen).

This example shows the retrieval of the Vβ-Jβ repertoire of 9 mice using parameters of Figure 20. The upper array displays the percentage of use of the 12 different profiles (here Vβ-Jβ) and in the lower array the fluorescence signal.

e) The 'Perturbation1' macro

This 'Perturbation1' macro is the first type of programmed analysis one can use to analyse perturbation of repertoires. This method implies the use of n=1 dimension distance. The method is based on an absolute value distance to compare each repertoire to an average control group (Gorochov, G., A.U. Neumann, A. Kereveur, C. Parizot, T.S. Li, C. Katlama,

M. Karmochkine, G. Raguin, B. Autran, and P. Debre. 1998. Perturbation of CD4+ and CD8+ T-Cell repertoires during progression to AIDS and regulation of the CD4+ repertoire during antiviral therapy. Nat. Med 4:215-221; Han, M., L. Harrison, P. Kehn, K. Stevenson, J. Currier, and M.A. Robinson. 1999. Invariant or highly conserved TCRα are expressed on double-negative (CD3+CD4-CD8-) and CD8+ T cells. J. Immunol. 163:301-311.). This distance is computed using the percentage of use of each CDR3 length in each profile.

Check that one has specified for each sample's group by its group number (Figure 20, column D). Select the 'Perturbation1' macro using the ISEApeaks menu bar. The user is asked to give the number of the control group. The control average repertoire is computed and checked: the sum of the usage of the peaks of each profile should be 100%. Then DataAnalyser calculates the distance between each repertoire and this average repertoire for each peak (Figure 23). Finally, the distance between a repertoire and the control average repertoire is calculated for each profile. ISEApeaks calculates the mean perturbation of a profile by averaging the perturbation of each profile. Status of the execution is indicated in the Excel status bar (bottom of the screen). Using the facilities of Excel® one can now plot and sort results. One can use the 'FillExcluded' macro to replace all occurrence of "nb_CTR=0" or "excluded" by the mean perturbation of the same profile for its group. This can be useful for subsequent statistical analyses such as MANOVA (Collette, A., S. Bagot, P.-A. Cazenave, A. Six, and S. Pied. 2002. A profound alteration of blood TCRB repertoire allows prediction of cerebral malaria. submitted.).

The upper array shows perturbation of each profile of each sample, lower array the perturbation of each peak compared to the average control repertoire ('Pc(Control)', column C].

f) The 'Perturbation2' macro

The 'Perturbation2' macro is another proposed method to assess the perturbation of repertoires that was described by C. Pannetier (Déchanet, J., P. Merville, A. Lim, C. Retière, V. Pitard, X. Lafarge, S. Michelson, C. Meric, M.M. Hallet, P. Kourilsky, L. Potaux, M. Bonneville, and J.F. Moreau. 1999. Implication of gd T cells in the human immune response to cytomegalovirus. J. Clin. Invest. 103:1437-1449.). This method also computes the distance between a repertoire and a computed average control repertoire but is different from the precedent one: the n-dimension distance used is the standard quadratic distance (n=2). Secondly, the data used is quantitative: CDR3 length analysis is performed but the percentage of use of each profile is also estimated (for instance Vα, to follow the example). This method can be extended to other quantitative data: the percentage of use of each profile can be estimated by FACS analysis (Déchanet, J., P. Merville, A. Lim, C. Retière, V. Pitard, X. Lafarge, S. Michelson, C. Meric, M.M. Hallet, P. Kourilsky, L. Potaux, M. Bonneville, and J.F. Moreau. 1999. Implication of gd T cells in the human immune response to cytomegalovirus. J. Clin. Invest. 103:1437-1449.), by semi-quantitative CDR3 length analysis (Even, J., A. Lim, I. Puisieux, L. Ferradini, P.Y. Dietrich, A. Toubert, T. Hercend, F. Triebel, C. Pannetier, and P. Kourilsky. 1995. T-cell repertoires in healthy and diseased human tissues analysed by T-cell receptor beta-chain CDR3 size determination: evidence for oligoclonal expansions in tumours and inflammatory diseases. Res. Immunol 146:65-80; David-Ameline, J., A. Lim, F. Davodeau, M.A. Peyrat, J.M. Berthelot, G. Semana, C. Pannetier, J. Gaschet, H. Vie, J. Even, and M. Bonneville. 1996. Selection of T cells reactive against autologous B lymphoblastoid cells during chronic rheumatoid arthritis. J. Immunol. 157:4697-4706;Caignard, A., P.Y. Dietrich, V. Morand, A. Lim, C. Pannetier, A.M. Leridant, T. Hercend, J. Even, P. Kourilsky, and F. Triebel. 1994. Evidence for T-cell clonal expansion in a patient with squamous cell carcinoma of the head and neck. Cancer Res. 54:1292-1297.) or with Taqman (Lang, R., K. Pfeffer, H. Wagner, and K. Heeg. 1997. A rapid method for semiquantitative analysis of the human Vb-repertoire using TaqManR PCR. J. Immunol. Methods 203:181-192.). For semi-quantitative CDR3 length analysis, DataFormatter will automatically calculate the percentage of use. For FACS analysis, the percentage should be put in the DataFormatter file, replacing the percentage calculated by DataFormatter.

The group numbers must be filled in the 'para' sheet file (Figure 20). Select the 'Perturbation2' macro using the ISEApeaks menu bar. The user is asked to give the number of the control group. Each DataFormatter file will be opened to retrieve the percentage of use of each profile. The control average repertoire is computed and checked: the sum of the usage of the peaks of each profile should be 100%. Then, DataAnalyser calculates the distance between a repertoire and this mean repertoire for each peak (Figure 24). Finally, the distance between a repertoire and the control average repertoire is calculated for each profile. Status of the execution is indicated in the Excel status bar (bottom of the screen).

The upper array shows perturbation of each profile and of the sample (row 2), the lower array shows the perturbation of each peak.

g) The 'RIS' macro

Another calculation has been used to described oligoclonality: the Relative Index of Stimulation (Cochet, M., C. Pannetier, A. Régnault, S. Darche, C. Leclerc, and P. Kourilsky. 1992. Molecular detection and in vivo analysis of the specific T cell response to a protein antigen. Eur. J. Immunol 22:2639-2647; Cibotti, R., J.P. Cabaniols, C. Pannetier, C. Delarbre, I. Vergnon, J.M. Kanellopoulos, and P. Kourilsky. 1994. Public and private V beta T cell receptor repertoires against hen egg white lysozyme (HEL) in nontransgenic versus HEL transgenic mice. J. Exp. Med. 180:861-872.). This analysis has been implemented in the ISEApeaks package.

Select the 'RIS' macro in the ISEApeaks menu bar. Indicate the group of samples to use as the control. To specify a particular sample, just give a new group identification to this sample. The average control repertoire is computed. RIS is calculated for each peak of each sample (Figure 25). The average RIS of each group is computed for each peak. Status of the execution is indicated in the Excel status bar (bottom of the screen).

The Relative Index of Stimulation (RIS) is calculated for each peak. Pc(Control) is the average computed repertoire. 'excluded' signals a profile that could not be analysed in this sample while '∞' signals a peak not present in the average control repertoire.

h) The 'OligoScore' macro

To assess the oligoclonality of a set of repertoires or to find recurrent peaks, a score was devised to quantify oligoclonality for each peak (Collette, A., S. Bagot, P.-A. Cazenave, A. Six, and S. Pied. 2002. A profound alteration of blood TCRB repertoire allows prediction of cerebral malaria. submitted., Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.). Launch the macro by choosing the 'OligoScore' item in the ISEApeaks menu bar. A score is first calculated for each peak, afterwards a global score is computed for each group (Figure 26, column M). One will find the recurrent peaks by sorting peaks using the scores of a particular group of samples compared to other groups (Collette, A., and A. Six. 2002. ISEApeaks: an Excel platform for GeneScan and Immunoscope data retrieval, management and analysis. Bioinformatics 18:329-330.). Status of the execution is indicated in the Excel status bar (bottom of the screen).

The upper array summarises peak numbers per profile for each sample. The lower array first shows the score per individual and the score per group (columns M and N). Peaks have been sorted according to the oligoclonal score of the first group, the second being the control (see Figure 20).

i) The 'Shohei' macro

Another useful utility has been implemented in DataAnalyser: One can import the peaks obtained for a quantification of a particular set of cells using a given Vβ, Jβ and CDR3 length (Hori, S. and al., manuscript in preparation). To do that, use DataParameter to make a parameter file that will order the n repetition of the same amplification in the position 1 to n. Then use DataExtractor and DataSmoother to extract and smooth the peaks. DataFormatter will then enable one to import the peaks into Excel. Check the peaks obtained and make the necessary correction.

Open DataAnalyser. Click the 'Shohei' macro in the ISEApeaks menu bar to start the analysis. Select the DataFormatter file where peak data are stored and indicate the correct worksheet. DataAnalyser then imports the peaks area and makes the calculations to estimate the number of cells using these particular Vβ, Jβ and CDR3 length.

j) The 'DrawArray' macro

Visualisation of repertoire diversity is not handy when analysing several repertoires. Colour-coding of diversity is a common way to visualise diversity (Pannetier, C., M. Cochet, S. Darche, A. Casrouge, M. Zöller, and P. Kourilsky. 1993. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor b chains vary as a function of the recombined germ-line segments. Proc. Natl. Acad. Sci. USA 90:4319-4323., Lim, A., A. Toubert, C. Pannetier, M. Dougados, D. Charron, P. Kourilsky, and J. Even. 1996. Spread Of Clonal T-Cell Expansions In Rheumatoid Arthritis Patients. Human Immunology 48:77-83; Mempel, M., B. Flageul, F. Suarez, C. Ronet, L. Dubertret, P. Kourilsky, G. Gachelin, and P. Musette. 2000. Comparison of the T cell patterns in leprous and cutaneous sarcoid granulomas - Presence of V alpha 24-invariant natural killer T cells in T-cell-reactive leprosy together with a highly biased T cell receptor V alpha repertoire. American Journal of Pathology 157:509-523; Musette, P., H. Bachelez, B. Flageul, C. Delarbre, P. Kourilsky, L. Dubertret, and G. Gachelin. 1999. Immune-mediated destruction of melanocytes in halo nevi is associated with the local expansion of a limited number of T cell clones. J. Immunol. 162:1789-1794.). ISEApeaks can automatically generate this representation. This macro is very useful as it is now possible to use this representation with objective classification of diversity as the Gorochov or Déchanet scores.

First parameterise the colour coding in the 'para' sheet (Figure 20). Select an array of cells where the values (for instance Gorochov values) have been entered: the array selected must contain the profile description but not the sample description. Launch the 'DrawArray' procedure with the ISEApeaks menu bar to create the new graphic (Figure 27).

For each profile, the diversity/perturbation is coded by a colour. The difference between the samples of each group is obvious.

k) The 'PeakNb' macro

One can obtain the number of peaks in each profile by using the 'PeakNb' macro. This number can be indicative of the oligoclonality of a repertoire.

l) The 'RepertoireMean' macro

The 'RepertoireMean' macro calculates the average repertoire for each group as determined in the 'para' sheet. The number of sample considered is also mentioned. This method can be used to identify public expansions (Han, M., L. Harrison, P. Kehn, K. Stevenson, J. Currier, and M.A. Robinson. 1999. Invariant or highly conserved TCRα are expressed on double-negative (CD3+CD4-CD8-) and CD8+ T cells. J. Immunol. 163:301-311.).

m) The 'CDR3Mean' macro

The 'CDR3Mean' macro calculates the average CDR3 length for each profile and for each sample. This can be useful as shown in Mugnaini et al. (Mugnaini, E.N., T. Egeland, A.M. Syversen, A. Spurkland, and J.E. Brinchmann. 1999. Molecular analysis of the complementarity determining region 3 of the human T cell receptor beta chain. Establishment of a reference panel of CDR3 lengths from phytohaemagglutinin activated lymphocytes. J. Immunol. Methods 223:207-216).

n) The 'SeparatemDescription' macro

The 'SeparatemDescription' macro achieves the visual separation of peaks by grouping them according to their mDescription. This macro will be useful if one has reordered the peaks to fit one's own criteria. Just select the zone in which one wants to separate the peaks, select the 'SeparatemDescription' in the ISEApeaks menu bar. One will be asked to indicate the column number of the column containing the mDescription (or other informations...).

o) The 'Export_Peaks' macro

The 'Export_Peaks' macro generates an array containing the percentage of use for each peak in its profile. This array is more convenient to export the data from Excel to other software (including statistics software such as StatView).

p) The 'Inverse_Array' macro

The 'Inverse_Array' macro is a utility to inverse any Excel array. The macro should be called when the data containing worksheet is active, then just indicate the array one wants to transpose. The result will be put in a separate worksheet. Alternatively, one can use the copy special of Excel to transpose an array.

q) Conclusion

By combining these tests and utilities, one can gain a good view of the repertoires that one would not be possible by eye. Additional scoring methods are currently in development and will be implemented in future versions.

Example 3.8: DataUtilities

DataUtilities proposes utility to deal with DataFormatter file.

a) The 'RenameNature' macro

. This utility is used to modify the designation of each well. One must specify the different files to be modified in the 'para' sheet while in the 'correspondence' sheet one will put, in the 'in' raw, the string of character to search for and, in the 'replace by' raw, the string of character one wants to use instead. Note that the 'NewName' raw is not used in this macro. After choosing 'RenameNature' in the ISEApeaks menu bar, select one of the DataFormatter files to process (all must be in the same folder). Then, in all the specified files, each occurrence of a 'in' character string will be replaced automatically by the corresponding 'replace by' character string.

In the DataUtilities Excel template, one can find worksheets with examples of parameterisation that have been used in the laboratory.

b) The 'ChangeOnemTheoricLength' macro

This macro will replace the mTheoricLength parameter in all sheets and all DataFormatter files indicated in the para sheet of a profile according to the mDescription and the new mTheoricLength values the user gave.

c) The 'ShitfmLengths' macro

This macro will shift the mLength value of all peaks in all sheets and all DataFormatter files indicated in the para sheet of a profile according to the mDescription and the new mTheoricLength values the user gave.

d) The 'DFConverter' macro

The 'DFConverter' macro allows the conversion of old DataFormatter files to the latest version of DataFormatter. Open DataUtilities, and fill the 'para' sheet: put the names of files to convert as well as the new name under which one wants ISEApeaks to save one's converted data (the sheet row is not used in this macro; therefore, a DataFormatter should appear only once). The use of 'para' sheet will allow one to convert a big set of files. Choose the 'DFConverter' item in the ISEApeaks menu bar.

Indicate the folder where one's old DataFormatter files are stored (the new DataFormatter files will be also saved in this folder). The conversion starts: each old file is open, data are transferred and percentage calculations are done, copying also the page set up.

e) The 'DFSpliter' macro

This macro allows one to separate data stored in DataFormatter files in separate data sheets of newly created DataFormatter files. For instance, imagine that one analyses the Vβ08.1-Jβ and Vβ08.2-Jβ repertoires and set the ISEApeaks add-in preferences in a way that these two groups of profiles are gathered in a common data sheet (12+12 = 24 profiles). DFSpliter will enable one to split the data in different data sheets in all the workbooks one precises in the 'para' sheet, like for the other DataUtilities macros. Fill the 'para' sheet, with the names of the workbook to split, indicate a template DataFormatter file (that should contain a data.1 sheet) and how many time a data sheet should be split (2 for the example). The macro will open each workbook and split the data. The macro checks that the number of profiles is dividable by the split number and that the preferences are suitable for the created workbooks (for the example, it should be 12).

Example 3.9: Examples

For each example, a 'Readme.txt' file will give one more detail of what to do to have a complete tour of the example.

DP & Parameter files

Files in the 'DP & Parameter files' folder provide CGEL parameter files and the corresponding Excel DataParameter sheets for 3SETCP v2.0 and 3SET2.1 v2.0: 36-well gel, 3 human Vβ-Cβ repertoires, m3SET v2.0: 36-well gel, 3 mouse Vβ-Cβ repertoires;

Vb2,3,5.1,4,16,7-Jb v2.0; and Vb6,9,8.1,8.2,14,8.3-Jb v2.0: analysis of 12 mouse Vβ-Jβ repertoire on two 36-well gels. These files can be used with DataParameter related macros.

DE, DS, DES & DF: Immunoscope data

Files in the 'IS Example' folder provide an example of a gel analysis from CGEL parameter file to the formatting of the extracted data in DataFormatter files.

DE, DS, DES & DF: GeneScan data

Files in the 'GS Example' folder provide an example of analysis of 2 gels from CGEL parameter file to the formatting of the extracted data in DataFormatter files. See the Readme.txt in this 'GS Example'.

DA

The 'DA' folder gathers DataFormatter files used in the DataAnalyser file 'DA 2.0 ex' where the result of all DataAnalyser macros is stored.

Utilities

The '.Pict files assembling' folder gathers data and results of an assembling of different PICT files according to a CPIC file parameter ('CPictPlaces para v2.0').

Example 3.10: Pitfalls

Some errors are quite usual when manipulating the programs. Remember that string in ISEApeaks must not contain the ';' character. DataExtractor opens none of the '.pict' files. Check the value of 'mTypicalPictFileName (Figure 17): is it correct? The name of the data files can be anything provided that it does not contain anymore figures than the 2 automatically assigned by the Immunoscope© software. The program will warn one if it is not the case. Also check that the '.pict' files CGEL parameter file one selected are in the same folder. Be careful not to confuse the different types of file used in the ISEApeaks package. Types and corresponding icons have been created to limit possibility of confusions. A good way to differentiate the different file types is to add a prefix to file names: 'Is m3SET' will be the Immunoscope macro created using the CGEL parameter file 'm3SET'. When using the CreateCGelFile macro, it might happen that Excel seems to be frozen. this is because other applications with whom the ISEApeaks Add-in is talking to do not signal to Excel that the job is finished. To avoid this, just after running CreateCGelFile, click on the desk of one's Mac (or any window not related to Excel). When Excel has finished, the Finder application becomes active. Just go back to Excel.

The following Examples are provided as analyses using ISEApeaks software in order to confirm the good working thereof.

Example 4.1 Perturbation of the CD4$^+$ and CD8$^+$ T lymphocyte repertoires during Leishmania infection in a murine experimental model

Example 4.2: Experimental model:

This study was carried out with wild-derived PWK inbred mice which was selected as a new promising model for human *Leishmania major* infection since the clinical and biological features of PWK mice parallels those observed in humans. PWK mice were infected by *Leishmania major* in footpads. 7, 20 and 27 weeks after infection groups of 6 , mice were sacrificed. Lymphocyte cell suspensions were prepared from draining lymph nodes (G) and spleen (R) and sorted into CD4$^+$ and CD8$^+$ sub-populations. Six additional mice were sacrificed just after infection and serve as controls.

Total ARN was extracted and reverse-transcribed into cDNA. The Immunoscope technique was applied to describe the diversity of VβCβ repertoires.

Example 4.3: Question

What is the perturbation of T lymphocyte repertoires during infection?

Is the perturbation different earlier (7 weeks) or later (27 weeks) after infection?

Is the perturbation different depending on the organ?

Is the perturbation different depending on the CD4$^+$ or CD8$^+$ sub-population?

Example 4.4: Results

Results are presented in the following pages and Figures 29-125 recapitulating this analysis:

- ISEApeaks data were collected from Immunoscope TCRBV-BC repertoire data and assembled into four "DataAnalyser Peaks" databases:
  - i. Spleen CD8$^+$ samples 0, 7, 20 and 27 weeks after infection
  - ii. Spleen CD4$^+$ samples 0, 7, 20 and 27 weeks after infection
  - iii. Lymph node CD8$^+$ samples 0, 7, 20 and 27 weeks after infection
  - iv. Lymph node CD4$^+$ samples 0, 7, 20 and 27 weeks after infection
- "Gorochov perturbation" (G1%) tables were generated to represent the global variability for each Vβ-Cβ profile of every sample with regard to the average variability measured for the control group. "Gorochov perturbation" were then computed by ANOVA for determining whether groups are statistically different based on their repertoire diversity.
- "Oligoscore" tables were generated for each group indicating the possible presence of recurrent oligoclonal peaks indicative of recurrent clonal lymphocyte expansions within a group. Heuristically, the threshold value corresponds to the maximal

"Oligoscore" value of the control group for which no significant expansion is expected. Only significant expansion are shown here.

- "Expression level" tables (R%) were generated to indicate the level of expression each Vβ-Cβ profile of every sample. Note that the experimental condition used here are not quantitative so that this value is only indicative when a major change is observed.
- Synthetic graphs for 7 and 27 weeks were drawn to represent the "Gorochov perturbation" of Vβ-Cβ combinations for which the "Gorochov perturbation" is statistically than control mice.

Example 4.5: Conclusions

On the basis of the elements of analysis obtained by the implementation of the ISEApeaks strategy described in the statements of invention DI99-92 and DI02-48 and the corresponding patent filed, we can support the following conclusions:

- $CD8^+$ repertoires appear to be less perturbed than $CD4^+$ repertoire especially in the spleen.
- The few Vβ-Cβ combinations which are perturbed among R $CD8^+$ are found among 7 and 27 week groups. On the contrary, the Vβ-Cβ combinations which are perturbed among G $CD8^+$ are mostly found among the 7 week group when the perturbation is close to control groups in the 27 week group.
- For $CD4^+$ repertoire, many Vβ-Cβ combinations are perturbed in the spleen and the lymph nodes. Again and more strikingly, the Vβ-Cβ combinations which are perturbed among G $CD4^+$ are almost always higher among the 7 week group than the perturbation of the 27 week group.

In another murine experimental model of parasite infection (PWK mice infected by *Leishmania major*), these results thus show the power of the ISEApeaks strategy to discriminate between subtle lymphocyte repertoire perturbations which would not otherwise be detected by eye.

Example 5.1: Effect of the LACK protein on the repertoire of αβ T lymphocytes in a murine experimental model

Example 5.2: Experimental model

This study was carried out with seven-week-old female BALB/C mice. They received an injection at the level of footpads of the LACK protein, or the LACKp (156-173)peptide, with or without treatment by an anti-IL-2 antibody. Every group of mice includes 5 individuals. Five additional mice received an injection of DMEM buffer and serve as controls.

The draining lymph nodes were taken 16 hours after injection. Total ARN was extracted and reverse-transcribed into cDNA.

First, the Immunoscope technique was applied to describe the diversity of VαCα and VβCβ repertoire, for Vβ4, Vβ8.1, Vβ3, Vα2, Vα8 and Vα15 TCRBV-BC combinations. Second, the diversity of Vβ4-Jβ and Vβ8.1-Jβ was analyzed.

Example 5.3: Question

Is it possible to detect a perturbation of the αβ T lymphocyte diversity following immunization by LACK protein or $\text{LACKp}_{156\text{-}173}$ peptide?

Example 5.4: Results

Immunoscope data was analyzed implementing the ISEApeaks strategy (data extraction, data smoothing, data formatting, peak database construction, data analysis). Results are presented in the following pages and Figures 29-125 recapitulating this analysis:

- Tables "DataAnalyser DA" created with ISEApeaks to prepare the extraction of the data through the various gels. It is at this level that the distribution of samples in groups is indicated.
- "DataAnalyser Peaks" tables corresponding to the database of peaks obtained for all individuals of all groups.
- "Gorochov perturbation" tables representing the global variability for each Vα-Cα, Vβ-Cβ or Vβ-Jβ profile of every sample with regard to the average variability measured for the control group (group n°1 - DMEM injection).
- "Oligoscore" tables for each group indicating the possible presence of recurrent oligoclonal peaks indicative of recurrent clonal lymphocyte expansions within a group. Heuristically, the threshold value corresponds to the maximal "Oligoscore" value of the control group for which no significant expansion is expected.
- "Gorochov perturbation" tables normalized by the "Oligoscore" values: for each individual and each Vα-Cα/Vβ-Cβ/Vβ-Jβ combination, the value of "Gorochov perturbation" of the individual was normalized by the ratio of the maximum Oligoscore for the Vα-Cα/Vβ-Cβ/Vβ-Jβ combination of the group considered by the maximum Oligoscore for the Vα-Cα/Vβ-Cβ/Vβ-Jβ combination measured for the control group.

- "DrawArray" tables for the "Gorochov perturbation" values.
- "DrawArray" tables for the "Gorochov perturbation" values normalized by "Oligoscore".

Example 5.5: Conclusions

On the basis of the elements of analysis obtained by the implementation of the ISEApeaks strategy described in the statements of invention DI99-92 and DI02-48 and the corresponding patent filed, we can support the following conclusions:

- Whereas neither Vα-Cα repertoire nor Vβ4-Cβ repertoires show any perturbation by comparison to the control group, Vβ8.1-Cβ repertoires show a tendency to perturbation.
- The comparison of the perturbation of the Vβ4-Jβ combinations to the average perturbation of Vβ4-Jβ combinations within the control group shows no difference. It's the same for the representativeness of every peak.
- Concerning Vβ8.1-Jβ combinations, perturbations superior to those of the control group are observed for Vβ8.1-Jβ1.1, Vβ8.1-Jβ1.2, Vβ8.1-Jβ2.2, Vβ8.1-Jβ2.7 combinations and to a lesser extent for Vβ8.1-Jβ1.3 combination, and this for the four experimental groups. This perturbation is not correlated to the representativeness. On the other hand, when we normalize the Gorochov perturbation by Oligoscore, it appears an increase of the perturbation correlated to the appearance of oligoclonal peak expansion for the following combinations:
    - Vβ8.1-Jβ1.1, Vβ8.1-Jβ1.2, Vβ8.1-Jβ1.3, Vβ8.1-Jβ1.4 within the LACK and LACKp groups

- Vβ8.1-Jβ2.1, Vβ8.1-Jβ2.2, Vβ8.1-Jβ2.3, Vβ8.1-Jβ2.4 and Vβ8.1-Jβ2.7 within four experimental groups.
- Vβ8.1-Jβ1.5 within the LACKp peptide groups, with or without treatment by the anti-IL-2 antibody.

In another murine experimental model, these results thus show the power of the ISEApeaks strategy to discriminate between subtle lymphocyte repertoire perturbations which would not otherwise be detected by eye.

Example 6.1: Kinetic of perturbation of αβ T lymphocyte repertoires in a murine experimental model of cerebral malaria

Example 6.2: Experimental model

This study was carried out with 2-month-old B10.D2 mice. They were infected by intraperitoneal injection of 10^6 *Plasmodium berghei* ANKA clone 1.49L parasitized red blood cells. We constituted six groups of mice: J3, J4, J5 and J6 groups (3, 4, 5 and 6 days after infection, respectively). J3 group included 5 mice when J4, J5 and J6 groups included 10 mice. The TSP group included 20 infected individuals and was used to follow the onset of cerebral malaria (CM). Parasitemia was systematically assessed before sacrificing the mice to confirm infection. Five additional mice were not infected and served as controls (TN group). Blood, spleen and brain were collected for each individual. For the TSP group, collection was done on the onset of CM clinical signs as assessed by paralysis, deviation of the head, respiratory troubles. For the TN group, collection was done on day 14 of the experiment. Total RNA was extracted and reverse-transcribed into cDNA.

The Immunoscope technique was applied to describe the diversity of Vβ-Cβ repertoire, for all the 23 Vβ TCRBV-BC combinations.

Example 6.3: Question

What is the kinetic of Vβ-Cβ repertoire perturbation during *P.berghei* infection?

Example 6.4: Results

Immunoscope data was analyzed implementing the ISEApeaks strategy (data extraction, data smoothing, data formatting, peak database construction, data analysis). Results are presented in the following pages and Figures 29-125 recapitulating this analysis:

- "DataAnalyser DA" table created with ISEApeaks to prepare the extraction of the data through the various gels. It is at this level that the distribution of samples in groups is indicated.
- "DataAnalyser Peaks" table corresponding to the database of peaks obtained for all individuals of all groups (not printed).
- "Gorochov perturbation" table representing the global variability for each Vβ-Cβ profile of every sample with regard to the average variability measured for the control group (non-infected TN group) (not printed).
- "Oligoscore" tables for each group indicating the possible presence of recurrent oligoclonal peaks indicative of recurrent clonal lymphocyte expansions within a group. Heuristically, the threshold value corresponds to the maximal "Oligoscore" value of the control group for which no significant expansion is expected (not printed).
- "Gorochov perturbation" tables normalized by the "Oligoscore" values: for each individual and each Vβ-Cβ combination, the value of "Gorochov perturbation" of the

individual was normalized by the ratio of the maximum Oligoscore for the Vβ-Cβ combination of the group considered by the maximum Oligoscore for the Vβ-Cβ combination measured for the control group (not printed).

- "DrawArray" tables for the "Gorochov perturbation" values.
- "DrawArray" tables for the "Gorochov perturbation" values normalized by "Oligoscore".
- Despite the small size of some experimental groups, Gorochov perturbation data was analyzed for each Vβ-Cβ combination by analysis of variance in order to identify statistically significant differences between groups. In order to confirm the preliminary observations, another analysis of variance was performed based on three sets of individuals only: one set comprising controls, J3 and J4 groups; a second set conrresponding to J5 group; a third set comprinsing J6 and TSP groups.

Example 6.5: Conclusions

On the basis of the elements of analysis obtained by the implementation of the ISEApeaks strategy described in the statements of invention DI99-92 and DI02-48 and the corresponding patent filed, we can support the following conclusions:

- The overall results of this experiment confirms our previous observation that repertoire perturbation can be observed during CM as compared to non-infected controls.
- From this study it can be seen that the repertoire perturbation is progressive during infection. The repertoires are globally not different between TN, J3 and J4 groups, on the one hand, and between J6 and TSP on the other hand. J5 group appears intermediate.

- For some Vβ-Cβ combinations the shift of perturbation happens between J4 and J5; for others, it happens between J5 and J6. This is confirmed when looking at the data obtained from the analysis of variance of the three sets [TN, J3, J4], [J5] and [J6, TSP].

These results suggest the predictive/diagnostic value Vβ-Cβ repertoire perturbation data. They show the power of the ISEApeaks strategy to discriminate between subtle lymphocyte repertoire perturbations which would not otherwise be detected by eye.

Example 7.1: Effect of vaccination with irradiated Plamodium yoelii sporozoites on the repertoire of αβ T lymphocytes in a murine model of malaria

Example 7.2: Experimental model

This study was carried out with C57BL/6 mice. Mice were constituted into groups as follows:

- A group of mice was immunized three times with irradiated Plasmodium yoelii sporozoites in order to induce protection to P. yoelii infection (I3*).
- A group of mice was immunized three times with irradiated Plasmodium yoelii sporozoites in order to induce protection to P. yoelii infection and later challenged with infectious Plasmodium yoelii sporozoites (I3*S).
- A group of mice was challenged with infectious Plasmodium yoelii sporozoites (S).
- An additional group of unmanipulated mice served as controls (T).

One week after challenge, spleen (R) and liver (F) were collected and lymphocyte suspensions were prepared. Total ARN was extracted and reverse-transcribed into cDNA.

The Immunoscope technique was applied to describe the diversity of Vβ-Cβ repertoire, for all the 23 Vβ TCRBV-BC combinations.

Example 7.3: Question

Is it possible to detect differences between repertoire diversity perturbation between groups?

In particular, is there a difference during challenge when mice have immunized or not with irradiated sporozoites (S vs. I3*S groups)?

Is there a difference between organs (liver vs. spleen)?

Example 7.4: Results

Immunoscope data was analyzed implementing the ISEApeaks strategy (data extraction, data smoothing, data formatting, peak database construction, data analysis). Results are presented in the following pages and Figures 29-125 recapitulating this analysis:

- "DataAnalyser DA" table created with ISEApeaks to prepare the extraction of the data through the various gels. It is at this level that the distribution of samples in groups is indicated.
- "DataAnalyser Peaks" table corresponding to the database of peaks obtained for all individuals of all groups (not printed).
- "Gorochov perturbation" table representing the global variability for each Vβ-Cβ profile of every sample with regard to the average variability measured for the control group (non-infected TN group) (not printed).

- Average "Gorochov perturbation" graph comparing the perturbation between organs (R & F) and experimental groups (T, S, I3* & I3*S) for each Vβ-Cβ combination and on average.

- "Oligoscore" tables for each group indicating the possible presence of recurrent oligoclonal peaks indicative of recurrent clonal lymphocyte expansions within a group. Heuristically, the threshold value corresponds to the maximal "Oligoscore" value of the control group for which no significant expansion is expected.

- "DrawArray" tables for the "Gorochov perturbation" values (not printed).

- Gorochov perturbation data was analyzed for each Vβ-Cβ combination by analysis of variance (ANOVA) in order to identify statistically significant differences between groups.

- Principle Component Analysis followed by Discriminant Analysis was performed in order to determine how many groups can be distinguished based on peak percentage information. The plot of data according to the first five factors is shown.

Example 7.7: Conclusions

On the basis of the elements of analysis obtained by the implementation of the ISEApeaks strategy described in the statements of invention DI99-92 and DI02-48 and the corresponding patent filed, we can support the following conclusions:

As summarized in the attached table, statistical analysis of data shows differences of perturbation repertoire between organs (R vs. F) or between experimental groups (in particular I3*S vs. S). This study brought us to design decisional trees to help decision in analyzing the data (see attached document). PCA/DA analysis also provides evidence that experimental group+organ combinations can be distinguished on the basis of the Vβ-Cβ repertoire diversity information. It can be seen that factors 1, 2 & 3 can discriminate between experimental groups F3* vs. F3*S vs. FT/RT/R3*S vs. FS/R3*/RS. This analysis will help determining the protective components of the immune response following vaccination by irradiated sporozoites.

In another murine experimental model, these results thus show the power of the ISEApeaks strategy to discriminate between subtle lymphocyte repertoire perturbations which would not otherwise be detected by eye.

REFERENCE TO COMPUTER LISTING OF RAW DATA AND ISEApeaks SOFTWARE APPENDICES

Filed herewith in triplicate (labeled Copy 2.1, Copy 2.2, Copy 2.3, Copy 3.1, Copy 3.2, and Copy 3.3 respectively) are computer listings of raw data on two separate compact discs read only memory (CD-ROM). The entire contents of the raw data appendix is incorporated herein by reference.

Each of the three copies of the computer listings of raw data appendix were created on July 1, 2003.

Filed herewith in triplicate (labeled Copy 4.1, Copy 4.2, and 4.3 respectively) is a computer listing of the software program ISEApeaks 2.0.1A on separate compact discs read only memory (CD-ROM). The entire contents of the computer listings of the software program ISEApeaks is incorporated herein by reference.

Each of the three copies of the computer listings of raw data appendix were created on July 1, 2003.

Numerous modifications and variations on the present invention are possible in light of the above teachings. It is, therefore, to be understood that within the scope of the accompanying claims, the invention may be practiced otherwise than as specifically described herein.